Fourth Year Project

Disaggregation of Rainfall Time Series using Adjustments

K. M. Wong

Supervisors: Dr. C. J. Onof & Prof. D. Koutsoyiannis

21st June 2000

This report consists of one paper and five appendices.

Declaration

This submission is my own work. Any quotation, or description of, the work of others is acknowledged herein by reference to the sources, whether published or unpublished.

____

Table of Contents

1. Introduction

Rainfall disaggregation is emerging as an important tool for both hydrologists and engineers to understand the hydrological processes that occur in nature. This knowledge is essential for civil engineering works, such as hydro-electric dams, flood alleviation schemes and the management of water-catchments. In the design of such schemes, rainfall data of both daily and hourly time scales are often required. However, there is usually a lack of sub-daily information, simply due to the absence of hourly raingauges. In most parts of the world, daily raingauges are far more prevalent than hourly raingauges. Even if such hourly raingauges do exist, they have usually only been in operation for a few years, making the length of the recorded series insufficient for making statistically significant conclusions.

Disaggregation concerns the generation of lower level time scale data (e.g. hourly rainfall data) that are consistent with higher level data (e.g. daily data). In this sense, disaggregation fills the knowledge gap described previously and allows hourly data to be generated even though no such historical records exist. Although these simulated time series may not be the actual rainfall depths that fell to the ground in the past, their statistics are consistent with the actual time series, as well as the higher time scale statistics. These statistics are often the most important data required in engineering design.

The objective of this paper to is examine whether hourly data can be obtained from daily data alone. Previously, most research on disaggregation has incorporated some form of low level statistic in its analysis. This paper examines the possibility of obtaining hourly time series using only daily or higher level statistics. This would have direct benefit on areas where daily data is freely available but hourly data is lacking.

The data set used in this paper comes from an hourly raingauge at Heathrow airport. There is access to both hourly and daily records for this site. Using a proven rainfall model together with a simple disaggregation procedure, hourly time series will be derived from daily statistics. By comparing the simulated hourly statistics to the actual recorded data, the effectiveness of the procedure can be determined.

The paper first begins by introducing the rainfall precipitation model used (section 3). The method used to fit the model's parameters and its corresponding computer programme are described in section 4. The disaggregation procedure and its programme are presented in section 5. The core of the paper occurs in section 6, where the analysis of different combinations of higher level statistics is examined. Detailed analysis of the various optimisation schemes is provided and their effectiveness is discussed. Reference to a new method of optimisation, the evolutionary algorithm, is also made in this section. Some examination of the extreme statistics of the simulated time series is given in section 7. This is especially important for determining the return periods of extreme rainfall events. Section 8 presents power spectrum analysis, which provides an insight into the underlying structure to the rainfall. The paper ends with a summary of the main conclusions and some suggestions for further research.

2. Notation and Definitions

Throughout this paper, many similar terms and statistics will be used and compared with one another, therefore it is imperative that these terms are formally defined so as to ensure clarity.

Firstly, **statistics** are defined as the actual characteristics of the rainfall time series, such as the mean, variance or autocorrelation. These are sometimes referred to as **data**. These two terms are distinct from **parameters**, which refer strictly to the parameters governing the Bartlett-Lewis point process.

High level statistics refer to those statistics of 24-hr and 48-hr time scales. **Low level** statistics refer primarily to 1-hr statistics, although in certain cases that are explicitly expressed, they can refer to 6-hr and 12-hr statistics. Coarse and fine statistics may also be used in place of high and low level respectively. **Autocorrelation** and **autocovariance** refers to the lag one autocorrelation and autocovariances.

Historical statistics are those which have been taken from historical records, they are actual data taken from rain gauges located on site. **Modelled parameters** are those that have been derived from these historical statistics, in this case using the method of moments. These modelled parameters are then used to obtain the **modelled statistics**, that is, they are statistics that are based on *derived* parameters. Finally, these modelled statistics must be distinguished from **simulated statistics**, which in turn refer to statistics derived after disaggregation via simulation. Both the modelled and simulated statistics are also called the **estimated statistics**, since they are essentially estimates of the historical statistics.

3. The Bartlett-Lewis Rectangular Pulse Model (BLRPM)

The rainfall model used in this paper is the Bartlett-Lewis Rectangular Pulse Model (BLRPM). This is a continuous time rainfall model for a fixed point in space. This model has been used with considerable success for a wide variety of climates, including the U.K. (Onof and Wheater, 1993, 1994). The model has proven useful for reproducing the statistics of both daily and sub-daily time scales (Rodriguez-Iturbe et al., 1987, 1988).

The BLRPM is a model for point rainfall time series modelling rainfall at a point, as opposed to spatial General Circulation Models which model rainfall over an entire area. The model incorporates Poisson cluster processes. Storms arrive according to a Poisson distribution and are represented by clusters of cells or rectangular pulses with constant depth. Fig. 3.1 shows the process more clearly. Each cell has a cell length and depth that is distributed exponentially, and it is the clusters of such cells that constitute the storms.

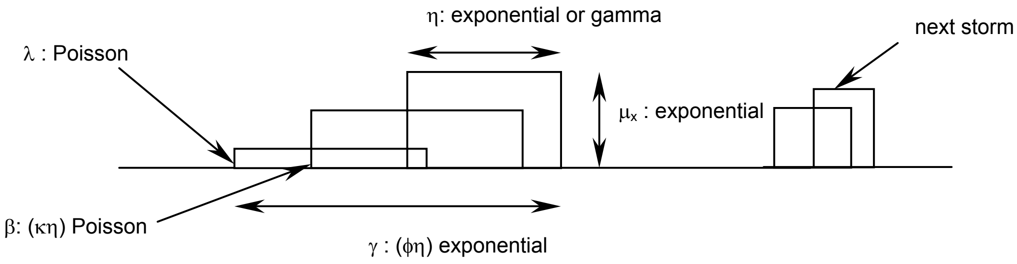


Figure 3.1 : Schematic of cluster process and parameters

The cluster processes within each storm can be represented in several ways. The Neyman-Scott process obtains the number of cells in each storm from a random distribution and uses an exponential distribution to obtain the cell arrival times (Rodriguez-Iturbe et al., 1987). The Bartlett-Lewis process, however, obtains the total storm duration from an exponential distribution, the cell arrivals being Poisson distributed. This process is described in further detail in the following sections.

3. 1 The BLRPM Five Parameter model

The original model outlined by Rodriguez-Iturbe et al. (1987) contained five parameters governing the storm and cell characteristics. Firstly, the parameter λ determines the storm arrivals according to a Poisson process. The parameter β determines the cell arrivals, also according to a Poisson process. The cell depths are governed by an exponential distribution with parameter μ_x. The entire length of the storm is determined by the parameter γ that is also exponentially distributed. Finally the individual cell lengths are distributed exponentially with parameter η. These are all shown in fig. 3.1.

For mathematical convenience, the parameters γ and β are often non-dimensionalised using the following equations:

$$\kappa = \beta / \eta \qquad \phi = \gamma / \eta$$

Therefore, for this model there are five parameters governing the process: λ, κ, μ_x, ϕ and η.

This basic model was found to perform well at various time scales, reproducing most of the statistics accurately. However, it was often found wanting in the fact that it could not reproduce accurately the proportion of dry periods for most time scales. The model generated fewer wet periods than required, underestimating the proportion of dry periods. In order to overcome this problem, Rodriguez-Iturbe et al. (1988) introduced an extra parameter into the model so as to give a better fit to the statistics.

3.2 The BLRPM Six Parameter Model

In this modified model, the cell length is no longer taken from an exponential distribution but is now determined by a gamma distribution with shape factor α and scale factor $1/\nu$. Hence, the parameter η is now replaced by two parameters: α and $1/\nu$. Therefore the model is now governed by six parameters: λ, μ_x, κ, ϕ, α and ν.

The statistics to be modelled can be calculated from the six parameters using the following equations:

Subcalculations:

$\mu_C = 1 + \frac{\kappa}{\phi}$ *(Mean number of cells per storm)*

$$k_1 = \left(2\lambda\mu_c\mu_x^2 + \frac{\lambda\mu_c\kappa\phi\mu_x^2}{\phi^2 - 1} \right)\left(\frac{\upsilon^\alpha}{\alpha - 1} \right)$$

$$k_2 = \left(\frac{\lambda\mu_c\kappa\mu_x^2}{\phi^2 - 1} \right)\left(\frac{\upsilon^\alpha}{\alpha - 1} \right)$$

Modelled statistics (equations 1-4):

$$Mean\ = \lambda\mu_x\mu_c \frac{\upsilon}{\alpha - 1} T \tag{1}$$

$$Variance = \frac{2\upsilon^{2-\alpha}T}{\alpha - 2}\left(k_1 - \frac{k_2}{\phi} \right) - \frac{2\tau^{3-\alpha}}{(\alpha - 2)(\alpha - 3)}\left(k_1 - \frac{k_2}{\phi^2} \right) \tag{2}$$

$$+ \frac{2}{(\alpha - 2)(\alpha - 3)}\left(k_1(T + \upsilon)^{3-\alpha} - \frac{k_2}{\phi^2}(\phi T + \upsilon)^{3-\alpha} \right)$$

$$Auto\,\mathrm{cov}\,ariance(lag - s) = \frac{k_1}{(\alpha - 2)(\alpha - 3)}\left\{ (T(s-1)+\upsilon)^{3-\alpha} + (T(s+1)+\upsilon)^{3-\alpha} - 2(Ts+\upsilon)^{3-\alpha} \right\}$$

$$+ \frac{k_2}{\phi^2(\alpha - 2)(\alpha - 3)}\left\{ 2(\phi Ts+\upsilon)^{3-\alpha} - (\phi T(s-1)+\upsilon)^{3-\alpha} - (\phi T(s+1)+\upsilon)^{3-\alpha} \right\} \tag{3}$$

$$\Pr(zero - rain) = \exp(-\lambda T - f_1 + f_2 + f_3) \tag{4}$$

$$f_1 = \frac{\lambda\upsilon}{\phi(\alpha - 1)}\left(1 + \phi\left(\kappa + \frac{\phi}{2} \right) - \frac{1}{4}\phi(\kappa + \phi)(\kappa + 2\phi) + \frac{\phi(\kappa + \phi)\left(4\kappa^2 + 27\kappa\phi + 36\phi^2\right)}{72} \right)$$

$$f_2 = \frac{\lambda\upsilon}{(\kappa + \phi)(\alpha - 1)}\left(1 - \kappa - \phi + \frac{3}{2}\phi\kappa + \phi^2 + \frac{\kappa^2}{2} \right)$$

$$f_3 = \frac{\lambda\upsilon}{(\kappa + \phi)(\alpha - 1)}\left(\frac{\upsilon}{\upsilon + T(\kappa + \phi)} \right)^{\alpha - 1}\frac{\kappa}{\phi}\left(1 - \kappa - \phi\frac{3}{2}\kappa\phi + \phi^2 + \frac{\kappa^2}{2} \right)$$

T is the respective time period (1, 6, 12 hour) and s is the lag period, usually taken as one.

Note that the equation for the proportion of dry periods is different from that quoted in Bo et al. (1994) or Rodriguez-Iturbe et al. (1988). There is a slight correction to a three of the ϕ terms in the expression of f_1.

The effect of adding this additional parameter is to randomise the cell length. The value of the storm cell length is no longer restricted to a pure exponential distribution, but rather occurs from a gamma distribution, of which the exponential distribution is but one form. This relaxation allows the model to be flexible in its representation of the rainstorms. The proportion of dry periods is now much better estimated using the six parameter model over the five parameter model and this is verified by Onof and Wheater (1993). However there are still some deficiencies, such as the underestimation of the skewness and the extreme values. The six parameter model was nevertheless used exclusively in the following analysis.

4. Parameter Fitting

4.1 Parameter Fitting: Method of Moments

There are several methods of fitting the BLRPM model to the historical statistics from Heathrow. The *Maximum Likelihood* method, although very commonly used in other models for parameter fitting, is unwieldy in this case (Rodriguez-Iturbe et al., 1987), as the likelihood function is difficult to obtain. The *Spectral Method* incorporating Fourier analysis can also be used. However, the *Method of Moments* is a better option, as it is significantly simpler and more practical to use. Also, the method of moments has been found to produce parameters that are significantly better than using spectral methods (Rodriguez-Iturbe et al., 1987).

The method of moments is set out simply as follows. More complex mathematical formulations of this method are set out in Rodriguez-Iturbe et al., 1988 and Cryer, 1986. The set of parameters to be fitted is given by the set Θ, where $\Theta = \{\lambda, \mu_x, \kappa, \phi, \eta\}$ for the five parameter model and $\Theta = \{\lambda, \mu_x, \kappa, \phi, \alpha, \nu\}$ for the six parameter case. Let k be the number of parameters to be fitted, either five or six. Next, p statistics are chosen from the historical data to fit the parameters, and these are denoted by the set T, where $T = \{t_1, t_2, t_3, \ldots\ t_p\}$. These can include the mean, variance, etc. of various time scales. The functions from which the various statistics can be calculated from the parameter values in the BLRPM (equations 1- 4) are denoted by the set S, where $S = \{s_1(\Theta), s_2(\Theta), s_3(\Theta), \ldots\ldots\ s_p(\Theta)\}$.

If $k = p$, then the method of moments requires:

$$S = T \qquad \forall\, p \tag{5}$$

This set of p equations is then solved for Θ, obtaining the parameter set. However, the functions that are within the set S are often highly non-linear, therefore it is difficult to obtain the Θ explicitly. Numerical methods must be used in these cases.

Often it is easier to formulate the equations in an error-residual form, such that the objective function is now:

$$\min \sum_{i=1}^{p} w_i \left(s_i(\theta) - t_i\right)^2 \qquad (6)$$

where w_i is the weight attributed to that particular statistic. These weights are usually set to unity in the following optimisation schemes. The aim is to find a set of parameters Θ where the expression in (6) is equal to zero, such that $s_i(\Theta)=t_i$ for all i={1, 2, 3, …..p}.

Therefore, this objective function is minimised in order to reduce the error between the calculated form of the statistics ($s_i(\Theta)$) and the actual historical data (t_i). In this formulation, p can be more than k. In some optimisation cases set forward later in the paper, p is less than k. This is not an ideal situation as we will have fewer equations than unknowns. Therefore, from a theoretical point of view, p must always be more than or equal to k so that there will be more equations than parameters to be estimated.

The choice of statistics to include in the set T is rather subjective. Ideally, these statistics should be independent and not highly mutually correlated (Rodriguez-Iturbe et al., 1987). Also the statistics must be able to incorporate all the parameters present in Θ. For most cases, the statistics chosen were the mean, variance, autocovariance and the proportion of dry periods corresponding to each of the individual time scales. The mean duration of dry periods may also be used, although it must be noted that this particular statistic incorporates data from two adjacent time scales.

4.2 Computer Programmes: `momentfit` and `Optima`

Computer programmes were used to implement the method of moments. `momentfit` is a FORTRAN coded programme, used on a UNIX platform. It uses the method of moments to fit parameters to the BLRPM, using the formulation set out in equation (6). `momentfit` uses deterministic optimisation algorithms from the NAG routine library to perform the parameter fitting.

This programme was deemed unsuitable for a number of reasons. Firstly, it required the one hour mean to be present in the parameter fitting; if absent, the optimisation routines within the NAG library would go into cycles and deteriorate. This requirement is clearly against the objective of obtaining parameters from daily data alone. Also, the programme was rather inflexible in its specification of parameter constraints. As the programme is located in a UNIX framework, subsequent data analysis was rather clumsy, since large amounts of data had to be repeatedly transferred from a UNIX platform to a Windows platform.

`Optima` was developed in order to allow parameter fitting to be performed in a Windows environment. This program is written in VBA (Visual Basic for Applications) and is located in the Microsoft Excel spreadsheet package, making it easier to manipulate and analyse data. The user enters the historical data via dialogue boxes and there is range of six different types of statistics (mean, variance, autocovariance, autocorrelation, proportion of dry periods and mean duration of dry periods) to choose from. The parameters are then fitted using the minimisation technique shown in section 4.2 above (equation 6). The SOLVER function is used to perform the optimisation of this objective function. Constraints to the parameters can also be easily modified. In addition, the user can also specify whether to fit parameters according to the six or five parameter model. Further specifications on these two programmes, as well as other computer programmes, can be found in the Appendix A.

The SOLVER function was used within Microsoft Excel to perform the optimisations. SOLVER uses the Generalised Reduced Gradient (GRG2) non-linear optimisation code developed by Leon Lasdon, University of Texas at Austin, and Allan Waren, Cleveland State University (1998). This method uses a deterministic, gradient based optimisation search scheme to find the optima. There are also some tests for optimality incorporated in the engine that allows the solver to test whether a global optimum has been reached. A different SOLVER engine was also tried, an evolutionary algorithm developed by Frontline Systems, Inc (1999). This is introduced in section 6.3.

`Optima` was subsequently used to perform all the optimisations cases. This was done after some testing to ensure that the results obtained under `optima` were similar to those produced by `momentfit`. Using `optima` allowed the use of pure daily data in the parameter fitting. Errors could be easily checked and corrected by changing the constraints or the initial values.

The optimisation path is rather sensitive to the initial values used to begin the optimisation. This is due to the highly complex nature of the objective function. The feasible region, being situated in six dimensional space, is liable to contain many local optima. The GRG2 solver uses gradient based techniques to find an optimum; therefore it may arrive at a local optimum rather than the global optimum. This requires many trials of different initial values, so as to ensure that the objective function value obtained is the actual optimum required. Often, the optimum was found to occur at point where one or more of the parameters was at a constraint. This meant that the optimum occurred at the edge of the feasible region. These observations were disregarded, since the constraints are set subjectively, and a better optimum could usually be found within the feasible region.

The evolutionary algorithm was introduced in the analysis to combat this problem of initial values. The initial point of optimisation is theoretically less of a problem when an evolutionary engine is used, since it randomly samples the feasible region for potential solutions. These ideas are developed further in section 6.3.

5. Disaggregation

5.1 Uses of Disaggregation

Disaggregation is an important step in the process to obtain lower-level time scale data from higher time scales. In this paper, the objective is to be able to model hourly (lower level time scales) using purely daily data (higher level time scales). Therefore, if we can obtain suitable parameters from daily data, we can use these parameters to perform a disaggregation and obtain hourly time series data.

In this project, the disaggregation was performed on the most promising sets of parameters. This was done as a check on the disaggregation procedure. Since disaggregation is a simulation, the statistics generated may fluctuate depending on the parameters used, as well as the random seed utilised in the simulation. Therefore, some variation from the historical and modelled statistics is to be expected.

Using the simulation also allows us to obtain statistics from the time series that are not explicitly modelled. Higher order statistics, such as the skewness, can be calculated from the simulated disaggregated time series, and this can be compared with the historical data to observe the goodness of fit. Disaggregated data can also be used to find out more about the extreme value plots of the simulated data. This will help determine whether the process can accurately estimate the return periods of the extreme events.

5.2 Procedure and methodology

Disaggregation refers to the method of obtaining lower time-scale time series and properties, from higher time-scale time series. For example, disaggregation can be used to derive monthly-level time series and statistics from an annual time series. Most conventional disaggregation models work by incorporating the lower-level variables within the higher-level variables using a linear function (Koutsoyiannis and Manetas, 1996). In this way, they reproduce the entire first and second order properties of the lower-level variables, as well as their mutual cross-correlations. The correlation between the higher and lower-level variables is also retained.

A simpler and more efficient method of disaggregation has been recently developed by Koutsoyiannis et al. (1996). This method is different from conventional methods in that it firsts generates a lower-level time series with no reference to the higher-level variables. There are essentially three steps to the process:

1. Generate a time series according to an appropriate model. In Koutsoyiannis' paper, the seasonal autoregressive (PAR(1)) model was used. In our case, the BLRPM was adjusted for use in the disaggregation.
2. Use an accurate adjusting procedure to correct the generated lower-level time series so that its terms add up to the corresponding higher-level variables.
3. Repeat the process until a suitable time series can be obtained which improves on the higher order statistics that are not explicitly preserved in Step 2

Step one generates the lower-level time series with no reference to the high level properties, simulating the rainfall according to the parameters entered by the user. Step two uses an accurate adjusting procedure to correct this time series so that the error between the sum of the generated lower-level series and the corresponding higher-level variable is reduced. These procedures are called accurate because they preserve certain statistics or in special cases, the entire distribution of the lower-level series even after the adjustment is performed.

There are three different methods of adjustment: the proportional, linear and power methods. The proportional method is ideal as it uses a very simple proportional scheme to correct the time series. However, it is only fully accurate for lower level variables with a gamma distribution incorporating a similar scale parameter as the higher level variables. The linear method has an advantage of being able to cope with any distribution, relaxing the constraints found in the proportional method. Its disadvantage is that it has a propensity to return negative values. The power method is a combination of the first two methods, and is able to perform calculations with the logarithms of statistics, but it is has the disadvantage of not being an exact procedure. This method returns only positive values.

The proportional adjusting procedure is used in this paper. It is mathematically expressed by:

$$X_s = \widetilde{X}_s \left(Z / \sum_{j=1}^{v} \widetilde{X}_j \right) \qquad (s = 1, \ldots, v) \qquad (7)$$

where $\widetilde{X}_s$ is a lower-level (e.g. hourly) synthetic series that has been generated by some stochastic model (in our case, the Bartlett-Lewis model). Z refers to a term in the higher-level data series Z_r (r = 1, 2, …). Therefore, the sum of v lower level time steps will equal to one higher level time step (r). X_s is the newly adjusted time series term.

Step three repeats the previous two steps until a time series is obtained that reproduces higher-order statistics (such as skewness) that resemble the actual data available. These statistics are not explicitly preserved in the previous adjusting procedure. This repetitive sampling may seem tedious and time wasting, but due to the parsimony parameters of the lower level model used, this method may be more efficient in the end.

For this paper, the BLRPM was used to generate rainfall at a low level time scale, using parameters optimised by various optimisation schemes. This corresponds to step one above of the disaggregation procedure. Adjustments were then made, so that the hourly rainfall depths added up to the daily depths. This was done using the proportional adjustment method. This process was then repeated until better values could be obtained for those statistics that were not explicitly preserved in the adjustment.

The coupling of the BLRPM and the adjustment procedure contains several problems. As the BLRPM is a continuous time model, as opposed to the discrete time disaggregation model, some assumptions and modifications have to be made. It is assumed that the clusters of wet days are independent (arrivals of the storms being a Poisson process), and therefore a separate BLRPM model is run for each cluster. This "discretises" the time series so that the adjustment procedure can proceed. Longer clusters of cells may have to be split into two or more cells for the model to cope.

5.3 Hyetos

The computer programme Hyetos was used to perform the disaggregation. This piece of software was produced by Koutsoyiannis and Onof (2000). In this program the user is required to enter in the parameters from the Bartlett-Lewis Rectangular Pulse Model; the six-parameter model is used. The actual historical rainfall time series can also be entered, so that the disaggregated and historical statistics can be compared. As an output, the programme gives the fully calculated statistics of the hourly time series, as well as the simulated time series obtained. Statistics are calculated for wet and dry periods as well as the whole time period.

Hyetos uses the proportionate adjusting procedure. As mentioned before, this procedure is exact given two conditions: the lower-level distribution must be a gamma distribution, and this gamma distribution must have the same scale factor as the higher level distribution of variables. This procedure was used partly because of its simplicity and also partly because it does not return negative values. Since rainfall time series contain many zero entries, if a linear adjusting procedure were to be used, the chance of obtaining negative values would be high. This will result in nonsensical results, as there is no such thing as a negative rainfall. Therefore, Hyetos implicitly assumes that rainfall is distributed according to a gamma distribution. Stationarity within each month is also assumed.

Repetition is also used in the programme to improve the statistics that are not explicitly reproduced by adjustment. In the process of repetition, a further refinement is made by considering the *distance*:

$$d = \left[\sum_{i=1}^{L} \ln \left(\frac{Z_i + c}{\tilde{Z}_i + c} \right)^2 \right]^{1/2} \qquad (8)$$

This term is the difference between the generated higher level terms and the actual higher level terms, and is calculated for every repetition of the time series. Once this distance reaches a minimum threshold level, the repetition stops and the adjusting procedure is applied. Therefore, this ensures that the generated results are close to the actual variables, thereby allowing accurate higher order statistics to be obtained.

Repetitions are user defined, and if the limit of repetitions has been met, the model will try to use other methods to model the rainstorm. If an especially long storm is encountered, the model will randomly split the storm in several portions, using a separate model to represent each portion. The programme then enters into a higher level of repetitions.

The procedure for the algorithm and the hierarchy of the various levels of repetition are shown more clearly by the chart in Appendix A, obtained from the help files within Hyetos.

The output of the programme can be in both text and graphical form. Graphs showing comparisons between the historical and simulated statistics are drawn up, detailing the skewness and proportion of wet periods, as well as the autocorrelation for lag *n* periods. More detailed statistics are given in text files.

6. Optimisation and Analysis

The data set consists of 39 years (1949-1987) of rainfall data from a single rain gauge at Heathrow Airport, Southwest London, U.K. The area is a wet region with approximately half the days of the year having some rainfall. The mean annual rainfall depth is about 600mm. Heathrow shows a rather stable climate, with the mean rainfall depths in the months of January and July being almost the same. These two months are classic "winter" and "summer" months, and the climatic stability is evident from the similarity of their rainfall depths.

The raingauge at Heathrow is an hourly gauge, therefore rainfall statistics at time scales hourly and higher were available. This allowed the effectiveness of disaggregation to be evaluated.

Data was analysed using the Fortran computer program `gaugestats`, which gave the various historical statistics for the time periods under analysis. Statistics were analysed individually by the month. For each month, parameters obtained from spectral analysis were also available.

Parameter fitting using `Optima` was performed for a variety of cases. The objective of the procedures was to use daily (24-hour) statistics to obtain BLRPM parameters, so as to predict smaller time-scale statistics (1-, 6-, and 12-hour). This was done in two steps: first obtain the BLRPM parameters using method of moments for a variety of different cases and initial parameter values, and then analyse each set of parameters to find their properties at smaller time-scales by using the BLRPM functions or disaggregation.

6.1 Four different cases: Optimisation using 1-hr, 6-hr, 12-hr and 24-hr statistics

To begin the study, historical statistics at daily and lower time scales were used. All optimisation was carried out using the method of moments, minimising the objective function described in equation 4 of section 4.2. All weights to the statistics were set to one. Four different cases were considered.

6.1.1 Description of four cases

Case 1 Only the 1-hour mean, variance, autocorrelation and proportion of dry periods statistics are used to obtain parameters. This is clearly not part of the objective, but is done as form of "control" in order to compare other results. However, it was discovered, as discussed later, that even the validity of using this case as a control was questionable. All parameters are initialised at 0.1 before the optimisation, this value being arbitrarily chosen.

Case 2 The parameters are first initialised at a value of 0.1. Next, the 1-hour mean, variance, autocorrelation and proportion of dry periods are used to get a set of parameters. This new set of parameters is then used as the initial values for the next optimisation. The next optimisation is now performed with the 6-hour mean, a total of five statistics being used in the optimisation. Another set of parameters is obtained and these are in turn used for the next optimisation. The next optimisation now includes the 6-hour variance, now there are a total of six statistics in all: 1-hour mean, variance, autocorrelation and proportion dry periods, and 6-hour mean and variance. This process is repeated, adding in the 6-hour statistics one by one, until all the 6-hour statistics are included (a total of eight statistics being used: 1-hour and 6-hour mean, variance, autocorrelation and proportion of dry periods).

Next, the 1-hour proportion of dry periods is taken out, leaving the optimisation to be done with seven statistics. This is repeatedly performed until all the 1-hour statistics are removed, leaving only the 6-hour statistics. Then the process is starts again, adding the 12-hour statistics one by one as before, and then removing the 6-hour statistics one by one. Finally, the 24-hour statistics are added and the 12-hour statistics removed. After a total 27 optimisations a final set of parameters is obtained. The diagram below gives a schematic of the process:

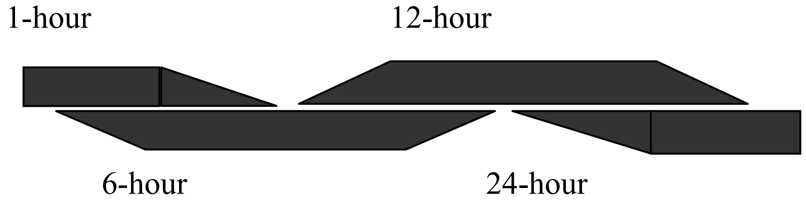


Figure 6.1: Schematic of optimisation process for Case 2

The bars represent the number of statistics being used in each optimisation, a full bar representing four statistics. On average, about 6 statistics are used to estimate 6 parameters, which is desirable.

Case 3 The parameters are first initialised at a value of 0.1. Next, the 1-hour mean, variance, autocorrelation and proportion of dry periods are used to get a set of parameters. This new set of parameters is then used as the initial values for the next optimisation. The next optimisation is now performed with the 24-hour mean, a total of five statistics used in the optimisation. A new set of parameters is obtained and these are in turn used for the next optimisation. The next optimisation now includes the 24-hour variance, now there are a total of six statistics in all: 1-hour mean, variance, autocorrelation and proportion dry periods, and 24-hour mean and variance. This process is repeated, adding in the 24-hour statistics one by one, until all the 24-hour statistics are included (a total of eight statistics being used: 1-hour and 24-hour mean, variance, autocorrelation and proportion of dry periods).   Next, the 1-hour statistics are phased out one by one until we are left with purely 24-hour data and a final set of BL parameters. This is a highly condensed version of Case 2, skipping out the steps using 6- and 12-hour data.

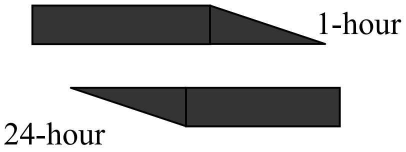


Figure 6.2: Schematic showing optimisation process for Case 3

Case 2 and 3 basically use 1-hour data to get an initial point of reference (essentially a Case 1 optimisation), before slowly adding in more statistics from larger time-scales for optimisation. Case 2 and 3 still require 1-hour data, therefore they do not entirely fulfil the objectives.

Case 4 Only 24-hour data are used in the optimisation, thus fulfilling the aims of the exercise. A variety of initial points are chosen and the objective function is minimised to obtain the optimal parameter set.

In summary, case 2, 3 and 4 are all basically optimisations that are based only on four

24-hr statistics. Case 2 and 3 are differentiated from case 4 in that their initial values are influenced in some way by statistics from lower level time scales. For case 2, numerous optimisations are performed, phasing in and out 1-hr, 6-hr and 12-hr statistics, so that a tentatively better initial parameter set is obtained to perform the final optimisation. For case 3, only the 1-hr and 24-hr statistics are used. For case 4, a range of arbitrary initial values is selected. Only this final case is truly independent of statistics from lower time scales.

The rationale for choosing these cases is based on the need to examine the sensitivity of the optimisation to the initial values. Cases 2 and 3 have their initial values influenced by the lower time scale statistics before the final optimisation using pure daily statistics is carried out. Case 2 was chosen to test the effect of 1-hr, 6-hr and 12-hr statistics on the initial values and subsequent optimisation, as well as to look for trends in the parameters during the phasing in and out of low level statistics. Case 3 was chosen to observe how hourly statistics could influence the initial values and the optimisation path. These cases are then compared to case 4 to determine if there is a significant effect of using initial values influenced by low level statistics on the optimisation process.

6.1.2 Analysis of the four cases

Trends within Case 2

This case was performed on the months of April, June and September.

This case was introduced in the hope that by phasing in and out the different statistics from different time scales, some sort of trend might be observed in the model parameters λ, ν, α, etc. The parameter values were plotted on to a histogram for each optimisation so that any trends could be identified, and a sample for September is attached below in fig. 6.3. Histograms from other months are placed in the Appendix B. As can be seen clearly from the graphs, there appears to be no clear-cut trend among the different parameters as the historical statistics from higher time scales are introduced in the optimisation. All the parameters seem to vary widely and there is no fixed pattern. When the histograms are compared to other months' graphs, there does not seem to be any underlying pattern. This observation does not prove that such a trend does not exist: if a different objective function, or another optimisation scheme, is used a trend may very well be uncovered.

Analysis prior to disaggregation

After each final optimisation was performed, the modelled parameters were substituted back into the equations 1-4 found in section 3.2, using *T* as one, six or twelve. These parameters were used to find the statistics for lower time scales. The results are the *modelled statistics*. These values are then compared with the historical statistics for goodness of fit. In this way, we can discover whether by using parameters optimised purely from 24-hr data, we can obtain accurate lower time scale statistics.

Table 6.1 shows the resulting BLRPM parameters for the four cases, while table 6.2 shows the modelled and historic statistics for the month of September. The shaded cells show the historic data, while the clear cells show the modelled statistics. Tables for April and June are also given in the Appendix B.

Table 6.1: BLRPM parameters for the four cases for September

λ	μx	κ	ν	α	ϕ	
0.013182	2.044199	0.343042	0.754222	3.406974	0.044585	Case 1
0.023331	2.426034	1.21067	0.022293	2.585572	0.013381	Case 2
0.019561	2.108476	0.521216	0.540097	3.2211	0.082485	Case 3
0.021472	28.74444	1.298096	0.402423	9.023899	0.945574	Case 4

Table 6.2: Historical and Modelled statistics for September

Case 1	1	6	12	24		Case 2	1	6	12	24
Mean	0.0734	0.4403	0.8807	1.7614		Mean	0.0734	0.4403	0.8807	1.7614
	0.073409	0.440453	0.880905	1.76181			0.072801	0.436805	0.87361	1.747221
Variance	0.196515	3.053057	8.118511	19.51673		Variance	0.196515	3.053057	8.118511	19.51673
	0.196513	3.016797	8.2047	21.41521			0.185646	3.131526	8.095259	19.51934
Autocor.	0.4877	0.3019	0.2051	0.1372		Autocor.	0.4877	0.3019	0.2051	0.1372
	0.487701	0.359836	0.305057	0.229231			0.583994	0.292542	0.205603	0.141847
Pro. dry	0.9225	0.8136	0.7208	0.5902		Pro. dry	0.9225	0.8136	0.7208	0.5902
	0.922505	0.848389	0.782827	0.668054			0.953127	0.848163	0.737365	0.557302

Case 3	1	6	12	24		Case 4	1	6	12	24
Mean	0.0734	0.4403	0.8807	1.7614		Mean	0.0734	0.4403	0.8807	1.7614
	0.073403	0.440417	0.880834	1.761669			0.073381	0.440284	0.880568	1.761136
Variance	0.196515	3.053057	8.118511	19.51673		Variance	0.196515	3.053057	8.118511	19.51673
	0.2056	3.101032	8.034782	19.51672			0.154968	2.966236	7.944915	19.51673
Autocor.	0.4877	0.3019	0.2051	0.1372		Autocor.	0.4877	0.3019	0.2051	0.1372
	0.488064	0.295501	0.214515	0.142712			0.672209	0.339225	0.228253	0.141909
Pro. dry	0.9225	0.8136	0.7208	0.5902		Pro. dry	0.9225	0.8136	0.7208	0.5902
	0.931092	0.840792	0.74754	0.591113			0.910538	0.828572	0.739274	0.588473

For the observations that follow, most of the examples are quoted from the statistics of September, although these trends were observed throughout the other months considered.

Case 1

Case 1 shows the worst results, with significant variation in the variances and autocorrelation. For example, although the 1-hr autocorrelation is very well estimated, the 6-hr, 12-hr and 24-hr autocorrelations are all overestimated, by up to 50%. It appears that parameters obtained using only 1-hr historical statistics are not suitable for the estimation of higher time scale statistics. This could indicate that the characteristics of the 1-hour data are not preserved when aggregation occurs. Other structures could be present and influence the higher time scales; there is no straightforward aggregation. For these reasons, case 1 must not be used as a "control"; all comparison must strictly be with the historical statistics.

Case 2

Case 2 gives better results than case 1, but still contains some statistics that are poorly estimated, such as the 1-hr autocorrelation. There was a similar trend for all the three months that case 2 was performed: at all times, case 3 proved a better optimisation scheme. There could be several reasons for this. Firstly, a greater number of historical statistics are used to find the initial starting points for the final optimisation, allowing a greater amount of error to be incurred. Any error in previous optimisations could be carried forward to the final optimisation, in the form of a poorly estimated model parameter set. This could result in a poor initial value for the optimisation, due to some error from previous optimisations. Also, the 6-hr and 12-hr historical statistics could include some non-stationarities or periodicities that are inherent in 6-hr or 12-hr time scales. For example, there could be some differences between a 12-hr nighttime period and a 12-hr daytime period. Such physical periodicities are less likely to occur in the more mutually regular 1-hr and 24-hr time scales. These are elaborated in the following section.

It was decided that case 2 was not a suitable optimisation scheme to proceed with, and therefore analysis using this case stopped after the three months aforementioned. Firstly, it is long-winded in its determination of the initial values and requires many levels of data. More importantly, it does not estimate the lower time scale statistics well. It has to be noted that this scheme may be improved if variable weights are added to the historical statistics used in the various optimisations, but this will generate more problems, such as the subjective question of what value of weight to assign.

Case 3

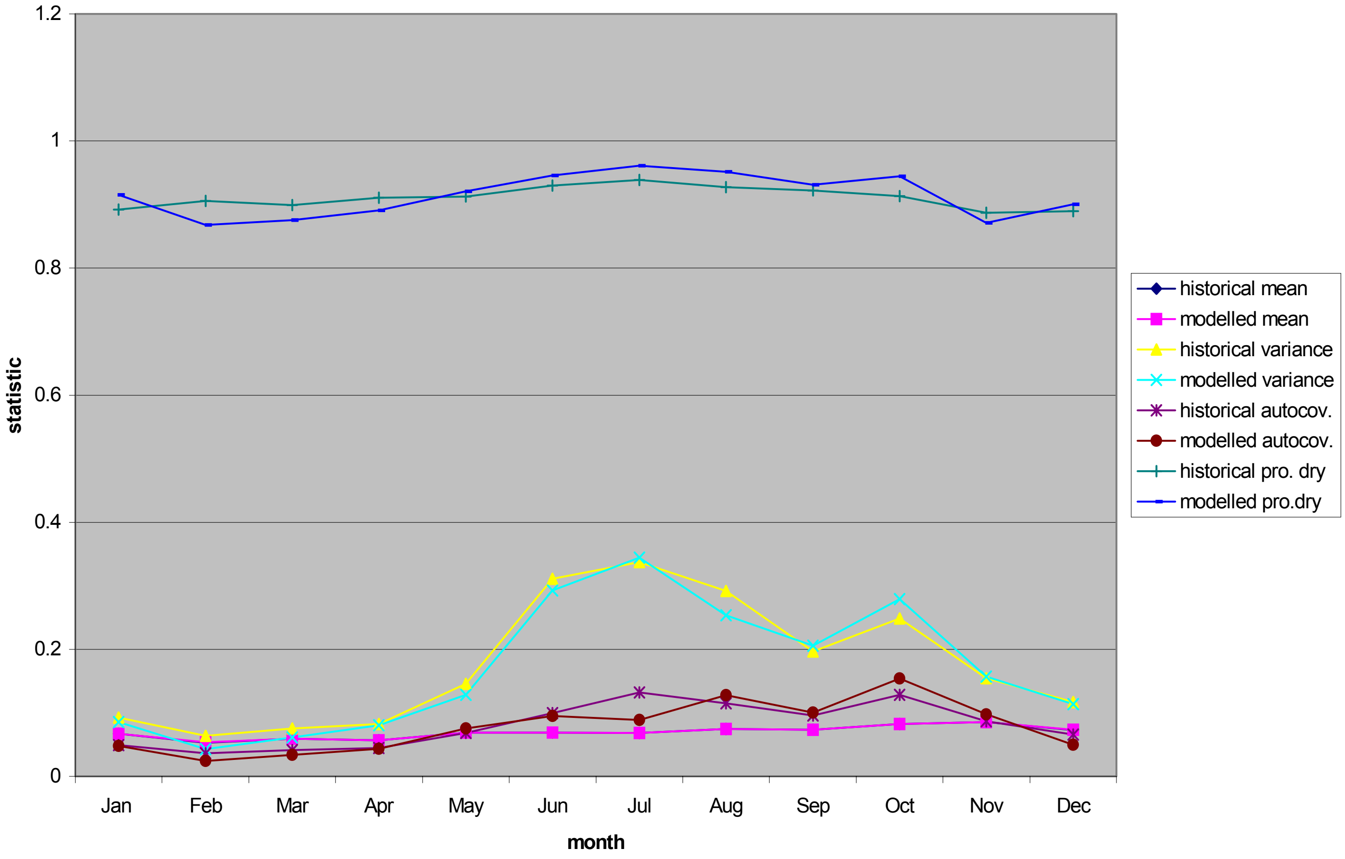
The optimisation scheme according to case 3 shows much better results in estimating the 1-hr statistics than case 2. In fact, case 3 can be considered to be the best case of the four. This in effect implies that the inclusion of 6-hr or 12-hr statistics does not help in obtaining accurate 1-hr modelled data. It is suspected that the cause of this arises due to the fact that the physical basis of 6- or 12-hr data has no meaning. 6 and 12 hours are simply arbitrary divisions of a half and quarter day used in statistical analysis. A 12hr night statistic is likely to be very different from a 12hr day statistic, likewise for the 6-hr time scale. This is not true, however, for the 1-hr and 24-hr periods, which show more regular and stationary behaviour. Therefore, using such statistics in the optimisation scheme may be of no use for obtaining accurate 1-hr modelled data.

Therefore, if hourly data is to be used to determine a suitable a initial starting point for an optimisation using only daily data, we may not need to go through 6- and 12-hourly data, but simply jump straight to 24-hour data. This proves to be the most accurate and efficient method of getting 1-hr modelled statistics from daily data. It must be noted that this scheme is not truly independent of hourly data, as it still uses some hourly data before the final optimisation.

Case 3 is better at estimating the 1-hr statistics than case 4. This shows that there is some effect in using 1-hr data to find a suitable initial starting point. Some characteristics of the 1-hr data have been preserved through the parameters and these influence the initial point of optimisation, allowing a better set of optimal parameters to be found.

Case 3 was repeated for every month for the data set and the results are shown in fig 6.4. As can be seen, the modelled statistics follow very closely the actual historic statistics. The mean and the autocovariance are very well estimated for almost every month. There are some discrepancies, with the variance having the most significant deviation from the historical values. This may be accounted by the fact that these months showed rather high objective functions in their optimisations.

Figure. 6.4: Case 3 historical and modelled statistics for the whole year



Case 4

Using purely 24-hour statistics gave surprisingly close results for all time scales. Using an example from Table 6.2, apart from the 1-hr variance, the other statistics are well reproduced. (Note that the error 1-hr autocorrelation is due to the error in the 1-hr variance alone, the 1-hr autocovariance is well estimated.) This case uses no hourly data to get the BLRPM parameters. This may mean that the daily data in fact incorporates the all the information all we need to get hourly statistics. Case 4 is still slightly less accurate than case 3, although it was found that for some months where case 3 did not give very accurate results, case 4 could obtain excellent results for the 1-hr modelled statistics (see Appendix C for October or November statistics). This is very encouraging as the findings imply that pure daily statistics may be used to accurately model hourly statistics.

The main problem of this optimisation scheme, and indeed all the following schemes, is finding a suitable initial point to start the optimisation. As proven by case 3, whose initial values are influenced by hourly statistics, a suitable initial point can give very good results. A wide range of initial points was used, with every parameter being initialised to this value, this being done for the sake of simplicity. Typically, the values of initial points ranged from 0.1 to 3.4. Higher values proved unsatisfactory, firstly because such high values are improbable as optimal BLRPM parameters, and also because such high values tended to skew the optimisation process toward giving a value of zero for the proportion of dry periods statistic.

A distinct characteristic of the optimisations is that a wide range of initial values often gave a single value for the objective function. For example, in table 6.3, the month of January gave an objective function value of about E-6 for optimisations with initial points 1.1, 1.25, 2.4 and 2.65. This could imply two alternatives: the presence of a local optimum in the feasible region of the optimisation, or the presence of a very flat region of the objective function. In the former case, the BLRPM parameters would be rather similar to each other at the end of the optimisation. In the latter case, the BLRPM parameter sets would differ, meaning that different points in the feasible region give the same value. The likelihood is that both phenomena are present.

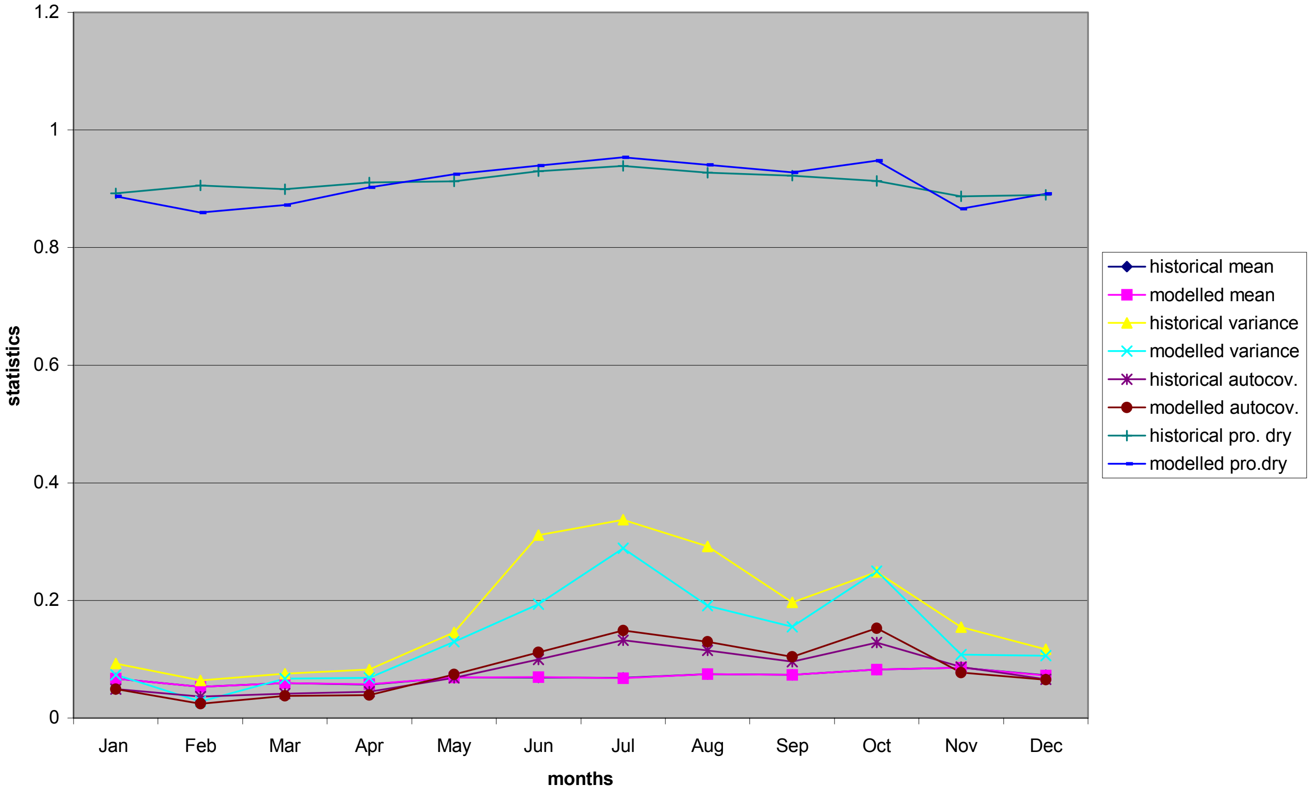
Table 6.3: Case 4 initial points and 1-hr modelled statistics for January

1-hr	historical	case 4 modelled statistics				
initial points	--	1.1	1.25	1.4	2.4	2.65
mean	0.0673	0.0673	0.0673	0.0628	0.0673	0.0672
variance	0.0927	0.0756	0.0695	0.1429	0.0806	0.0735
autocovariance	0.0495	0.0483	0.0484	0.0514	0.0481	0.0489
proportion of dry periods	0.8922	0.8913	0.8879	0.9206	0.8877	0.8871
objective function	--	E-6	E-6	0.09	E-5	E-6

Another observed characteristic is that, if such a local optimum exists for different initial points, it tends to be the lowest objective function value. Take for example January, as mentioned before, there exists a local optimum with objective function in the region of E-6. However, when optimisation begins with initial values of 1.4, the objective function converges on the value 0.09. This means that a different local optimum has been found. The presence of many local optima is a common trend, this can lead to confusing sets of parameters due to the complexity of the feasible region. The cause of this could be the fact that only four statistics are being used to estimate six parameters. The important observation is that the local optimum with the lowest objective function value tends to be returned from the most number of initial starting points for the majority of the months.

It would be natural to conclude that this lowest local optimum is the global optimum, and this may be true at least within the feasible region defined by the constraints. However, there is no direct evidence of this fact, due to the complex nature of the objective function. There is no optimality test for the global optimum. There may very well exist another optimum with an even lower objective function value. Therefore, this value and its corresponding set of BLRPM parameters can only be referred to as the lowest local optimum. By identifying the lowest local optimum for each month and using its corresponding set of BLRPM parameters to substitute into equations 1-4, we can obtain modelled 1-hr statistics. This procedure was performed on Case 4 for every month, using a wide range of initial points. These are the values plotted against the historical statistics in fig 6.5. In this way, the best parameter sets are selected for the range of initial points tested in each month, and used to derive the hourly modelled statistics.

Figure 6.5: Case 4 historical and modelled statistics for the whole year



The lowest local optimum point is important as it shows several distinct characteristics. As can be seen from fig. 6.5, the common trend among the months is that the 1-hr mean and autocovariance are very well estimated. The proportion of dry periods is also quite well estimated, although there is a slight deviation for the months of October and February. The variance is almost always underestimated, although for a few cases, it is estimated very accurately. It can be concluded that using this lowest local optimum from case 4 to estimate hourly statistics gives good estimates for the mean, autocovariance and the proportion of dry periods, and a reasonable estimate for the variance. It is also noted that the biggest discrepancies in the variance occur in the summer months.

Within each month, there are also several important observations. For those initial points that arrive at the lowest local optimum, the hourly mean, autocovariance and the proportion of dry periods tend to be estimated well, and if there are any deviation from the historical values, then the deviation tends to be consistent. For example, in table 6.3, the proportion of dry periods is very consistent at a level of 0.887, below the historical value of 0.8922. The hourly variance can have significant deviation, although it is invariably underestimated, and tends not to be as consistent. Considering only the optimisations arriving at the lowest local optimum in January, the variance can range from 0.069542 to 0.080551. These values are always below the historical value 0.09272. Other examples can be found in Appendix C.

For other local optima with high values of objective function, these trends are not followed. Results may arise with variances that are overestimated (almost never the case for the lowest local optimum) or autocovariance very badly estimated. However, it was also observed that there were some cases were the hourly statistics were extremely accurately estimated. For example, in April (table 6.4), there is a clear lowest local optimum for the initial points of 0.2, 0.3, 1.4 and perhaps 2.4. Initial values 0.8 and 2.65 give poor high objective function values. However, notice that the hourly statistics are accurate for 0.8, yet very poor for 2.65. This may be just a random occurrence that the BLRPM parameters are very good for this high objective function value.

Table 6.4: Case 4 initial points and 1-hr modelled statistics for April

1-hr	historical	case 4 modelled statistics					
initial points	--	0.2	0.3	0.8	1.4	2.4	2.65
mean	0.057	0.056965	0.057281	0.055558	0.056609	0.056506	0.058199
variance	0.082197	0.06225	0.068546	0.08239	0.054036	0.071847	0.048186
autocovariance	0.044493	0.039686	0.039172	0.046805	0.039718	0.044867	0.036641
proportion of dry periods	0.9104	0.911512	0.902163	0.93552	0.906066	0.888418	0.894033
objective function	--	E-5	0.0002	0.005	0.0008	0.0011	0.006

It must be noted at this point that the global optimum obtained under the optimisation using hourly data is not likely to be equal to the global optimum obtained under optimisation using daily data. There will be some error between the two parameter sets. Therefore, obtaining a very good fit for the daily statistics (low objective function) does not guarantee that there will be an equally good fit using the same parameters for the hourly statistics. Hence, there could arise the situation where an imperfect fit is found (such as another separate high-value local optimum) using daily data, yet the hourly data are modelled accurately. This could explain the existence of accurate hourly modelled statistics from high objective functions.

In summary for this case, it has been found that a wide number of optimisations starting from different initial points tend converge to a single value in the objective function. This value is usually the lowest found for all the trials performed. Termed the "lowest local optimum", the modelled statistics derived are found to fit closely with the historical data, with the exception of the variance, which is underestimated. Trials with initial points that do not arrive at this lowest local optimum have higher objective functions. The modelled statistics derived from these parameters are often poorly estimated when compared to historical data, although there are a few notable exceptions where the statistics are highly accurately estimated.

6.2 Cases using more high-level statistics

For each of the cases above, the final optimisation is performed with only four statistics. As mentioned before in section 4.2, this is clearly inadequate, as there are only four equations estimating six parameters. By adding more statistics from the 48-hr time scale, we can use six or eight statistics to find the optimal parameters. An example of the results can be found in table 6.5, the rest of the statistics being found in Appendix C.

Table 6.5: Various results and modelled statistics for November

1-hr	historical	8stats							
initial points	--	1.1	1.4	1.85	2.2	2.4	2.65	evo	evo
mean	0.0856	0.085829	0.085385	0.085675	0.085463	0.085327	0.08561	0.085602	0.085566
variance	0.154528	0.219853	0.109858	0.118516	0.090057	0.08888	0.161487	0.087652	0.10492
autocovariance	0.087076	0.080104	0.08238	0.084607	0.073497	0.073263	0.078628	0.072973	0.078894
proportion of dry periods	0.8868	0.894785	0.857938	0.839325	0.861613	0.851047	0.878991	0	0.085839
objective function value	--		0.4	0.4	0.4	0.4	0.4	0.69	0.4
The mean and the proportion of dry periods are well estimated. There may be two local minima (1.4/1.85 and 2.2/2.4) The evolutionary algorithm gives a higher or equal objective function									

1-hr	historical	6stats			
initial points		2.2	2.4	1.85	evo
mean	0.0856	0.084121	0.085343	0.086503	0.085463
variance	0.154528	0.100575	0.094522	0.08347	0.105903
autocovariance	0.087076	0.076525	0.074413	0.06589	0.079535
proportion of dry periods	0.8868	0.877323	0.875732	0.871254	0.857328
objective function value		0.4	0.4	0.4	0.4
Very similar to 8stats, since the error in the objective function lies primarily in the daily variance. The evolutionary algorithm manages to obtain the same local minimum as the GRG2 engine, although it takes a much longer time					

6.2.1 Optimisation using eight statistics (8stats)

For this case, eight statistics are used, namely the 24-hr and 48-hr mean, variance, autocovariance and proportion of dry periods. These eight statistics were used to optimise the parameters. The results are shown in fig 6.6 on the following page.

The general trend is similar to that of Case 4. Once again, the main problem in the optimisation is to find a suitable initial starting point. Once a range of starting points have been tried, local optima must be identified to find the best-estimated set of BLRPM parameters. Most of the feasible initial points converged to a lowest local optimum, while a few rare cases converged to other higher optima. These latter cases usually gave poorly modelled statistics.

An interesting observation is that the initial values that gave the lowest local optimum in case 4 were not guaranteed to be the same initial points that produced the lowest local optimum under optimisation with eight statistics. This meant that the whole range of initial starting points had to be tested, and the best initial points under case 4 could not be used as initial points for this optimisation.

Another observation is that the range of feasible initial values is much less than before. In all cases, initial points of values less than 1.1 were not valid, because the optimisation path tended to bring the parameter values to one of the lower constraints. If the initial values were set to less than 1.1, the optimisation will cause the μ_x value to fall to its lower constraint, resulting in an abnormally high objective function. This was also the case for high initial values. This time, the optimisation path tended to cause the parameter κ or ν to veer to their upper limits, causing the objective function to be too high.

As can been seen from fig. 6.6, the estimated hourly statistics are not as good as those under case 3or case 4. For this optimisation scheme, the objective function tended to be higher than previous cases. This is due to larger number of statistics being used in the optimisation, which leads to more difficulty in obtaining a set of parameters that gives low errors to every statistic used. Objective function values were of the range

Figure 6.6: 8stats historical and modelled statistics for the whole year

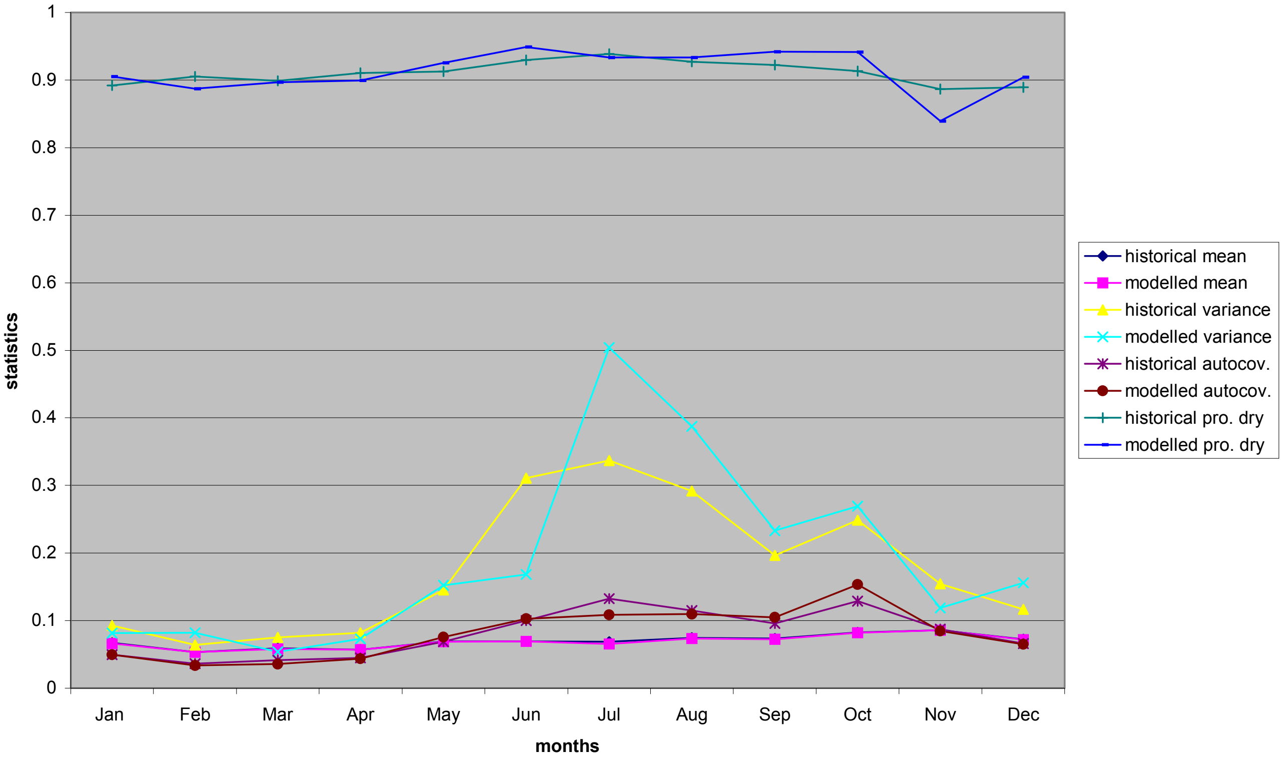
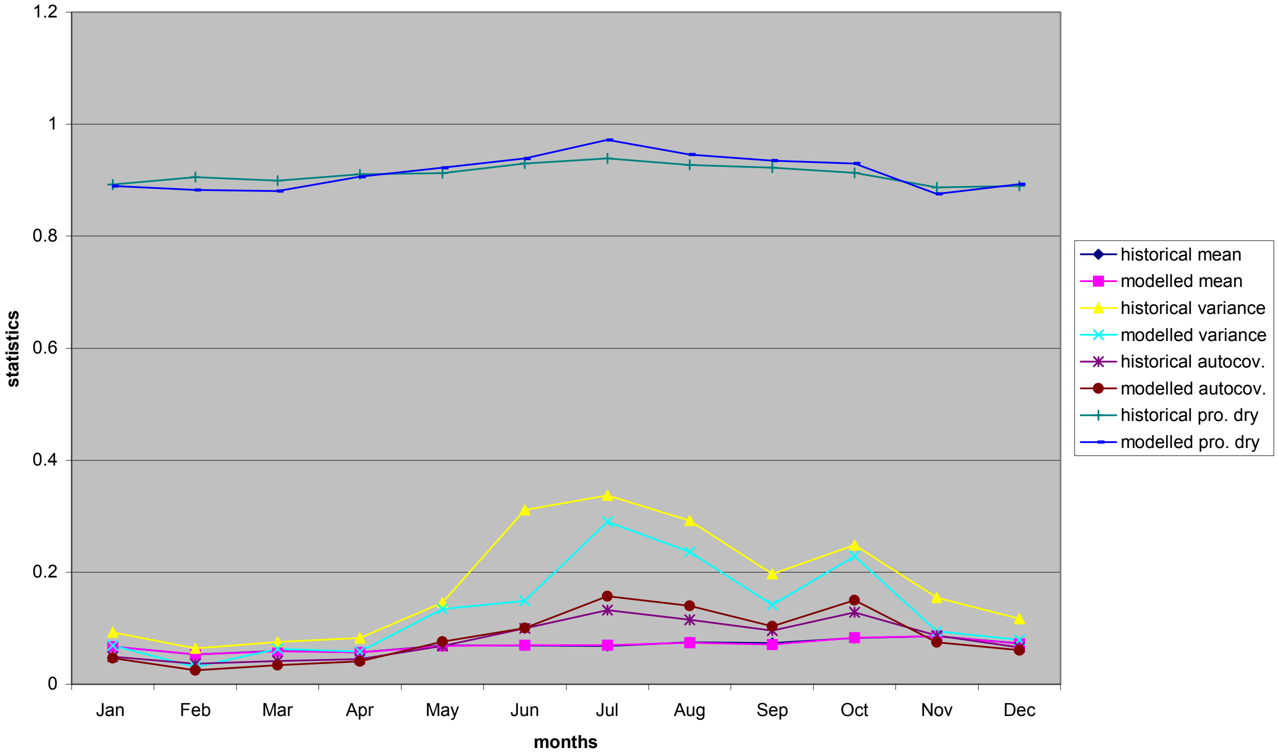


Figure 6.7: 6stats historical and modelled statistics for the whole year



E-3 to E-1, this is much higher than the typical range of E-7 to E-5 of case 4 (see Appendix C). This is also reflected in the plots, as the modelled statistic shows a much larger deviation from the historical values. Previously, the variances were almost always underestimated, whereas in this case, the variances are overestimated for quite a few months (February, July and October). For the summer months (June to September), the modelled variances are significantly inaccurate. This could be due to the fact that these months contain complex seasonal weather patterns that are difficult to model. It is noted that the objective function for July and August were extremely high in the order or 1.4 - 4.0, with most of the error arising from the daily variance term. This high objective function value could explain the inaccuracy of the estimated variance, suggesting that if a lower local optimum could be found, the variance may be better estimated.

The other statistics are relatively well estimated. The proportion of dry periods shows some deviation, especially in the month of November, when compared to other cases. This once again may be attributed to the high number of statistics being used to estimate the parameters, allowing more errors to be incurred.

6.2.2 Optimisation using six statistics (6stats)

Optimisation was performed using six historical statistics. These included the 24-hr mean, variance, autocovariance and proportion of dry periods, and the 48-hr variance and proportion of dry periods. The 48-hr variance was included in the hope that it would improve the estimation of the hourly variance, which seems to be the most difficult statistic to estimate. The 48-hr autocovariance was deemed to be well estimated and having low errors, hence it was not included in this optimisation scheme. The 48-hr mean is incorporated in the 24-hr mean, and therefore it does not require being included. This leaves six statistics to be used to estimate six parameters, a theoretically sound case.

The optimisation scheme was tried with a wide range of initial values, and the results are shown in fig. 6.7 before on page 32. Once again, the lowest local optimum was identified and these statistics were plotted. The lowest local optimum tended to be easier to identify than under the 8stats case. The conclusions set before under Case 4 and 8stats were generally found to be evident under this optimisation scheme.

The optimisation process tended to follow closely with the 8stats optimisation scheme. Most of the error incurred in the objective function under this scheme arose from the daily variance and autocovariance. Therefore, as the optimisation proceeded, the error accumulated at these two daily statistics, even as the other statistics were optimised to very accurate levels. Hence, on first appearance, the optimisation scheme did not seem to be very different from 8stats.

However, it is the modelled parameters and their corresponding statistics that are important and it was found that this optimisation scheme gave much better estimates of the summer months variances, which were poorly estimated in the 8stats scheme. The proportion of dry periods also did not show any irregularity. This could probably be accounted by the fact that the objective function only incorporates six statistics, leaving less room for error. Having an equal number of statistics and parameters to be estimated may have also had an effect on the improving variance estimates.

The graphs show that the estimated results under the 6stats case are in fact very similar to case 4. This implies that although theoretically unsound, from a purely empirical viewpoint, case 4 actually gives good estimates of the hourly statistics. Therefore, taking a pragmatic view, it may be feasible to use only four daily statistics to get hourly estimates of parameters. This will require less data to be gathered and less computation. However, as this method is not theoretically sound, it remains an empirical observation, and will need further verification using other data sets.

In summary for the cases of eight and six statistics, the conclusions that were presented for case 4 were invariably manifest in these two cases. This can been seen from the data tables in Appendix C. Once again a lowest local optimum was identifiable, and the estimated statistics from this parameter set were close to the historical data. The variance was once again an exception, being underestimated most significantly in the summer months. The objective functions for these values are higher than before, reflecting the fact that more statistics, and therefore more sources of error, are being used. Modelled statistics from 6stats are a much better estimation of the historical statistics than 8stats.

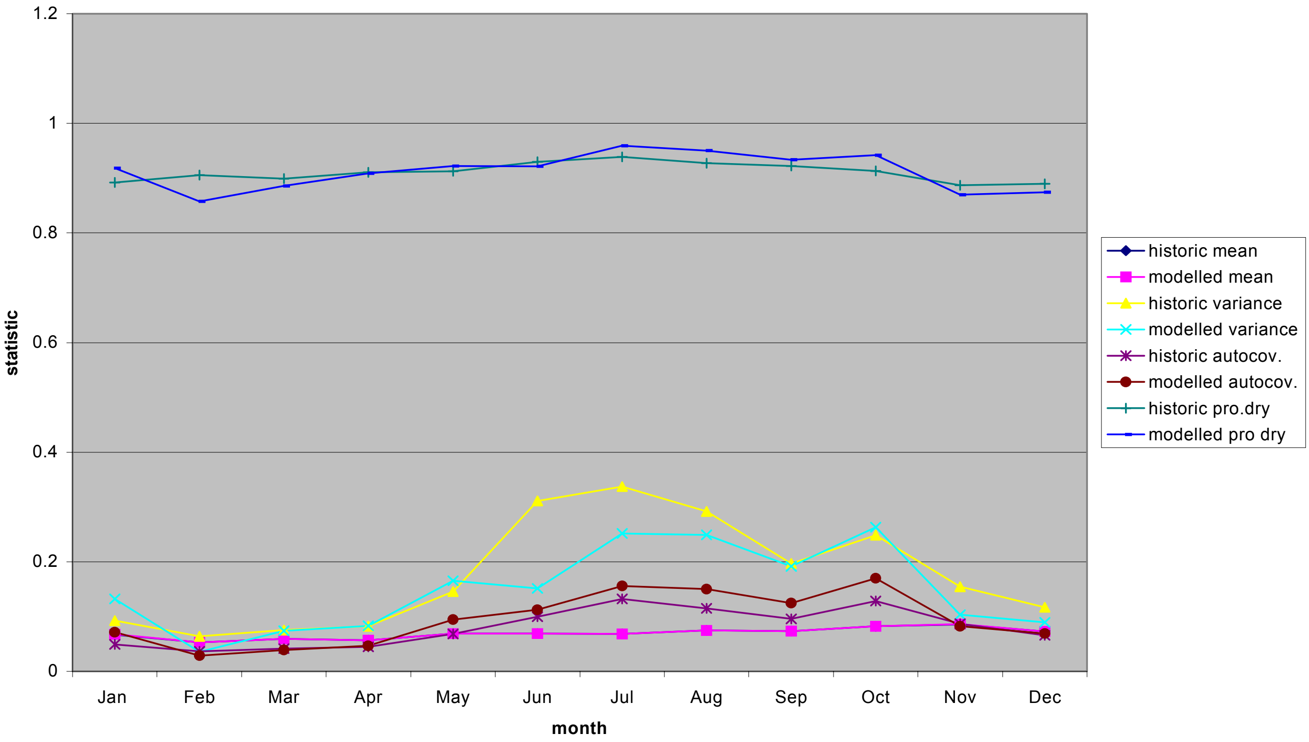
6.3 Disaggregation

Disaggregation is performed on the modelled parameters for two main reasons. The first reason is to obtain a disaggregated time series so that hourly, albeit simulated, data can be produced for practical engineering purposes. Disaggregation was also performed in the hope that such a process would improve the hourly modelled statistics, such that they more closely resemble the historical data.

6.3.1 Disaggregation using 6stats parameters

Hyetos was used to disaggregate the best sets of parameters for each month. Using the BLRPM parameters obtained under the 6stats optimisation scheme, Hyetos was run to obtain a simulated disaggregated time series. The parameters selected corresponded to the lowest local optimum obtained within the range of initial values used. The statistics corresponding to this disaggregated time series are then compared to the actual historical data. The results are shown in fig. 6.8 below.

Figure 6.8: Disaggregated 6-hr historical and modelled statistics



The results from the disaggregation follow very closely the results from using the modelled statistics. Hyetos is essentially a simulation, and therefore it is expected that the final results will resemble the modelled statistics, since the two processes are based on the same set of BLRPM parameters. However, there seems to be small improvement in the estimation of the statistics. The simulated statistics tend to lie closer to the modelled statistics rather than the historical statistics. Therefore, disaggregation fulfils the first objective set forth in section 6.3, but probably not the second. Any improvement is due to the repetition used within disaggregation to obtain better time series.

Nevertheless, disaggregation does provide a powerful tool for obtaining hourly time series data. Combined with the findings from section 6.2, disaggregation forms an empirical method of obtaining hourly time series and statistics from pure daily statistics.

6.3.2 Other disaggregation schemes

Modelled parameters from other optimisation schemes were also disaggregated. These were done on parameters derived under case 3, case 4, 8stats and 6stats, chosen for their excellent reproduction of modelled statistics, such that they closely resembled the historical statistics. This was done to test the process of disaggregation; to ensure that the disaggregation process does not alter the statistics. This is a different criterion than under section 6.3.1, where sets of parameters were chosen when their objective function was a lowest local optimum.

It was expected that the simulated statistics and time series after disaggregation would closely resemble the historical statistics, since these parameters were chosen precisely for the close fit between the modelled and the historical statistics. Some of the results are shown in the fig. 6.9-10 in the following pages, and in Appendix D. It is noted that the simulated statistics show highly accurate estimates of the historical data, for both the dry period statistics and the wet period statistics. The autocorrelation for lag periods of more than one are also closely estimated. In particular, the skewness was also very well simulated, showing rather accurate results,

Figure 6.9: Results from Hyetos for disaggregation with BLRPM parameters optimised under Case 4 (initial value of 0.3) for September

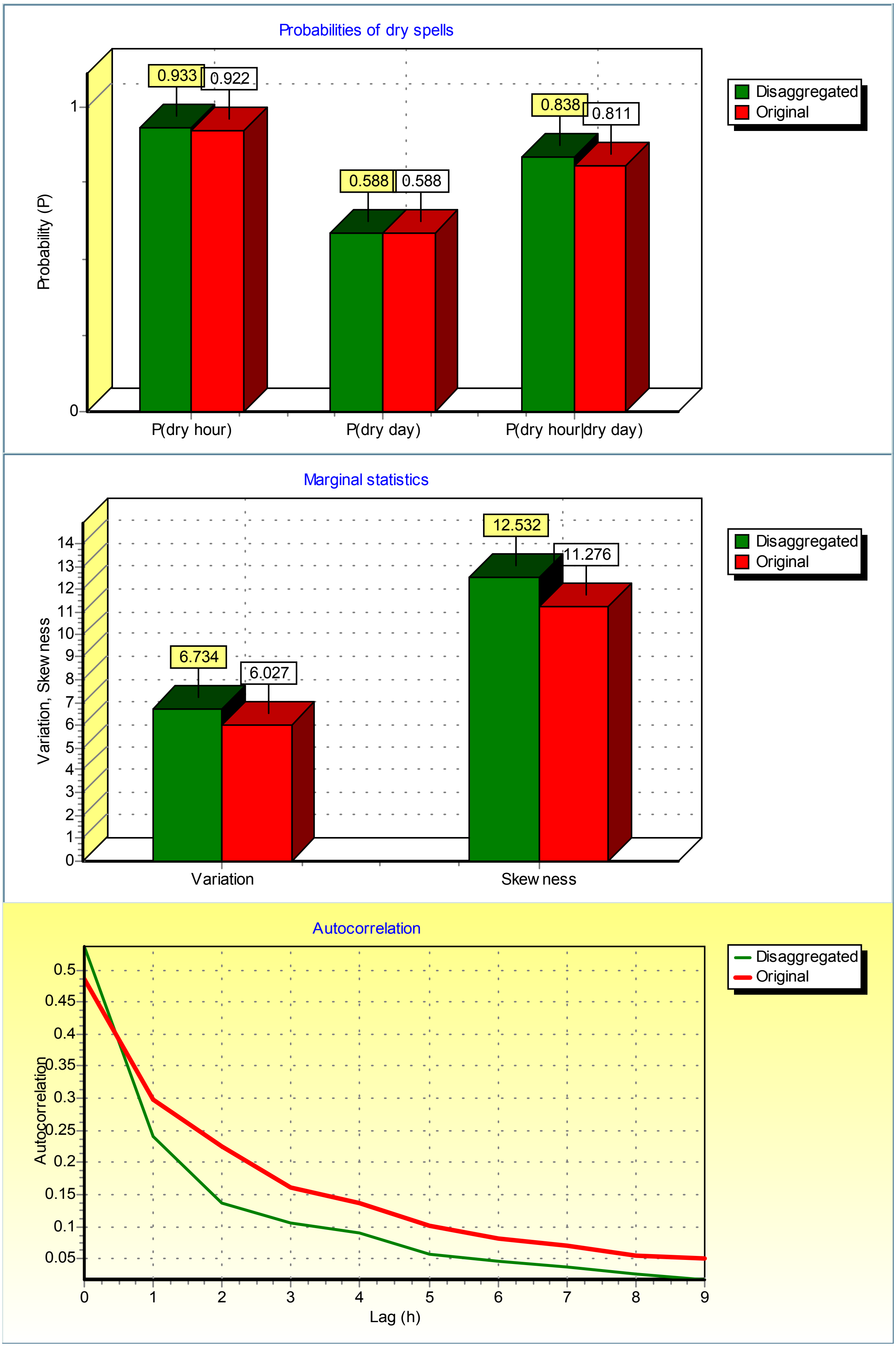
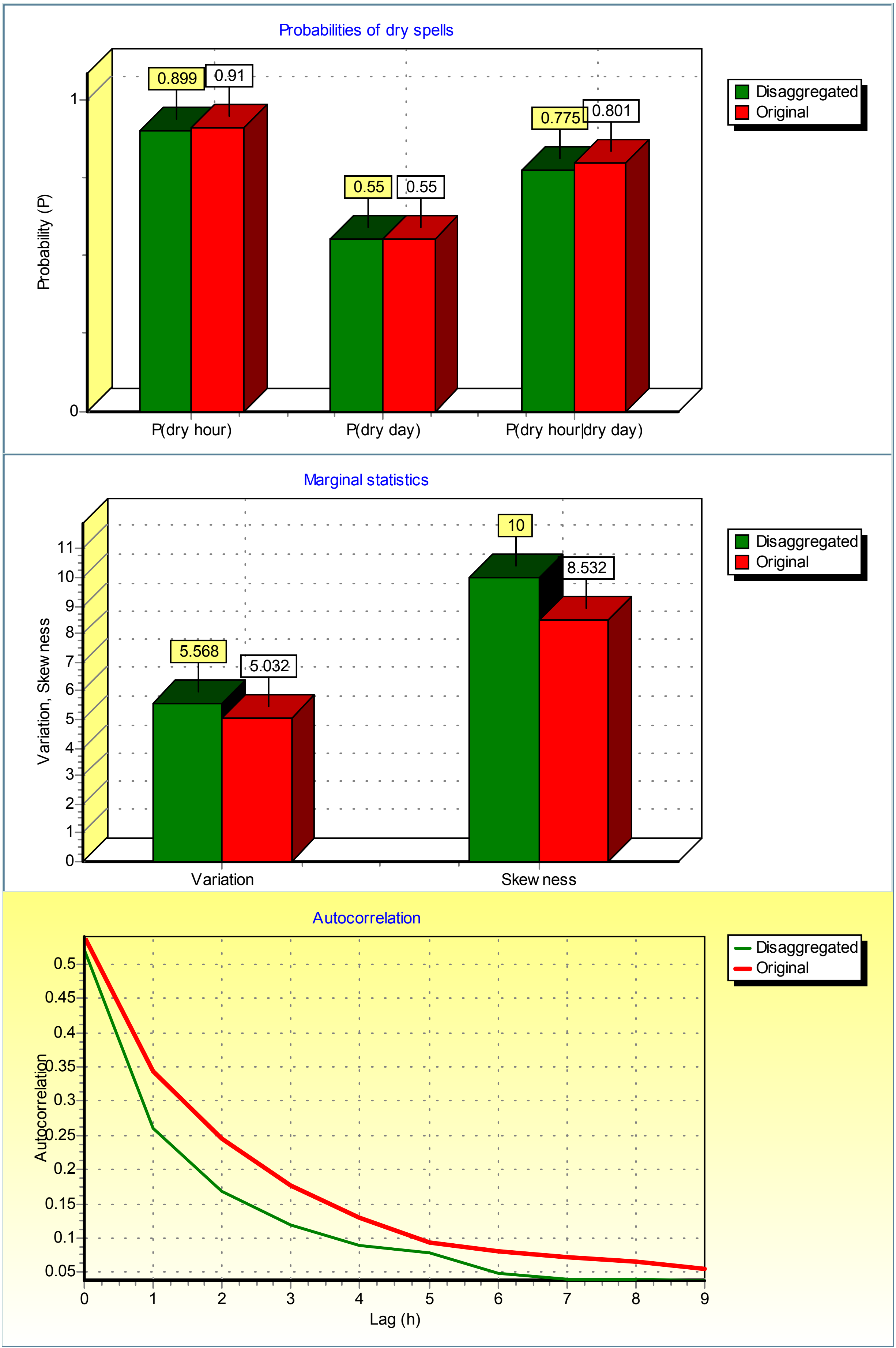


Figure 6.10: Results from Hyetos for disaggregation with BLRPM parameters optimised under Case 3 for April



even though this statistic was not explicitly modelled in the BLRPM. This is encouraging, as it is a third order statistic. Therefore, although disaggregation may not improve the modelled statistics, it may give accurate estimates of other higher order statistics such as skewness.

In summary, disaggregation gives a close fit with the modelled and historical data, hence if a good set of parameters can be found from the 24-hr or 48-hr data, the 1-hr statistics can be estimated by disaggregation to high accuracy. An hourly time series can therefore be derived from daily data. Other statistics, such as skewness and lag-n autocorrelations also can be estimated accurately, provided an excellent set of BLRPM parameters is used.

6.4 The Evolutionary Algorithm

In recent years, significant advances have been made in optimisation methods. One of the most prominent and unusual breakthroughs has been the development of genetic algorithms (GA). These are algorithms that are based on the principles behind the theory of evolution, such as survival of the fittest, mutation and reproduction. Such algorithms have been widely used for optimisation problems in a whole variety of fields. These algorithms are only just beginning to be implemented in the field of hydrology.

For this project, one of the major problems encountered was the selection of an appropriate initial starting point, as the deterministic nature of gradient-based optimisation algorithms causes the optimisation path to be sensitive to the initial points of the optimisation. Genetic algorithms have the unique characteristic of being non-deterministic, and therefore largely unaffected by the initial starting point. Therefore, if such algorithms can be effective in the optimisation for parameter fitting, there would be great benefit as there would be no need for a large number of trials to try out different initial points.

The Evolutionary Solver GA was used to perform the optimisations in this project. This is an evolutionary algorithm that has been developed by Frontline Systems, Inc., specially as a Microsoft Excel SOLVER engine. Therefore, by downloading the package from the website and installing it within the Excel spreadsheet application, the Evolutionary Solver can be used as an alternative optimisation engine. The new algorithm can be simply substituted for the original default SOLVER engine. In order to implement this, a variant of the programme `optima`, name `evo`, had to be developed to cope with the needs of the new SOLVER engine.

6.4.1 Basic principles of the Evolutionary Algorithm

The evolutionary algorithm's code and algorithm details are much too complex to be discussed in this paper, therefore only the basic principles behind the method will be presented. Much of the following material has been provided by the developers from Frontline Systems, Inc. (1999).

Random Sampling

Unlike normal gradient-based optimisation procedures, the evolutionary algorithm relies on random sampling. Solutions are randomly picked from the feasible region, as defined by the constraints explicitly set forth by the user. This allows the algorithm to explore large areas of the feasible region simultaneously. This therefore prevents the algorithm from being trapped within local minima; if one solution has reached a local minimum, there will still be other solutions available in other regions that may reach an even more optimal solution.

Populations

The solutions are kept together in a set called the population. This population is the essence of the algorithm, as it contains all the potential candidates for the optimum solution. In this respect, the algorithm differs from deterministic methods that keep only one potential solution throughout the optimisation process. The evolutionary algorithm does have a "best" solution within the population, but it also keeps the other solutions in case they are able to replace this "best" solution.

Mutation and Reproduction

The population mirrors nature by imposing on the set of solutions a mutation rate. This is a constant, user-defined rate at which a solution is extracted from the population and randomly changed, or "mutated". This is done in the hope that the mutated solution may be an improvement from its previous incarnation. Of course, there is an equal likelihood that the solution will be worse off, and if this is the case, then the algorithm will attempt to "repair" the mutation, though not with guaranteed success. This particular algorithm uses three different methods of mutating the solutions.

Two good solutions may be also combined together, forming a better solution. This is based on the process of reproduction in nature. The new solution will carry the best characteristics of both its "parents", and therefore it is hoped that it will be a better solution. This algorithm uses two methods of reproduction.

Natural Selection

Both the previous processes are performed in order to improve the solution set. If a more optimal solution is found, that solution will become the next "best" solution. At each iteration step (called a subproblem), the worst, or "least fit" solutions are removed from the population, being replaced by "more fit" solutions. This mirrors the principle of natural selection, or survival of the fittest. In this way, the population is constantly improved, with weak solutions being removed and strong solutions entering in.

Disadvantages

However, there are quite a few substantial drawbacks to the method. Firstly, the process is atheoretic, it merely an analogy to nature and does not have a solid theoretical background. Also, due to the large amount of calculation required at each subproblem, the algorithm is rather time consuming. It can take a few hours to perform the optimisation for the BLRPM parameters, while a normal gradient-based optimisation would take a few minutes at most.

The algorithm can only judge the strength of a particular solution in relation to the other solutions. Therefore, it can never actually know when a solution is truly optimal. There is no optimality test to perform on the solutions. Therefore, even after the algorithm has been completed, the optimal solution may still not be found, simply because the algorithm has not explored a particular region, and cannot comprehend that its present solution is not optimal. For this reason, the evolutionary algorithm should be used on optimisations with difficult or no tests for optimality, simply because there would be no better alternative. More importantly, because it has no test for optimality, it does not know when to stop. Heuristic methods have to be used to discontinue the algorithm; these include putting a limit on the convergence of the solutions in the population, and limits on time as well as iteration steps.

For this project, the optimisation of the BLRPM parameters is highly complex, and does not have a test for optimality. Therefore the optimisation has inherent difficulty in the judgement of "optimal" solutions. Hence, this disadvantage of the evolutionary algorithm should not significantly affect the optimisation scheme.

6.4.2 Analysis of results

The evolutionary algorithm was performed on five different months: January, April, August, September and November. The algorithm was used on case 4, 8stats and 6stats, all the optimisation schemes that previously required a range of initial points to be tested. The results can be seen in the data tables within Appendix C, where they are marked under the columns "evo".

Table 6.6: Various Results and modelled statistics for April

	historical	**8stats**							
initial point	-	0.1	0.2	0.3	0.8	1.5	2.65	evo	evo
mean	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.056	0.057
variance	0.0822	0.056	0.06	0.055	0.073	0.059	0.071	0.049	0.148
autocovariance	0.04449	0.041	0.041	0.041	0.044	0.041	0.041	0.041	0.056
proportion of dry periods	0.9104	0.908	0.909	0.9	0.899	0.906	0.905	0.86	0.929
objective function value	-	0.006	0.006	0.007	0.006	0.006	0.006	0.02	0.016
	historical	**6stats**							
initial point	-	0.4	0.8	1.4	1.8	2.4	2.65	evo	
mean	0.057	0.054	0.057	0.056	0.056	0.056	0.057	0.056	
variance	0.82197	0.089	0.058	0.075	0.072	0.05	0.086	0.06	
autocovariance	0.04449	0.041	0.041	0.042	0.041	0.04	0.041	0.042	
proportion of dry periods	0.9104	0.927	0.906	0.912	0.917	0.9	0.906	0.09	
objective function value	-	0.014	0.002	0.003	0.004	0.002	0.002	0.018	

Although the evolutionary algorithm was deemed to have an exceeding amount of potential, it was a failure in terms of practical implementation. This was a disappointment, as had it had succeeded in finding optimal solutions, it would have solved the problem of finding the appropriate initial points. Take for example the results in fig. 6.6, the objective function values obtained from the evolutionary algorithm are much higher than those of the GRG2 engine are. The corresponding modelled statistics are also poorly estimated.

The main problem with the algorithm engine was the length of time it took. Often, the optimisation was performed over night, but only mediocre results were obtained. The algorithm is effective at reducing the objective function value to the order of 1.0-4.0 at a fairly rapid length of time. However, once it enters into a refinement stage, the improvements in the objective function are few and far between. For example, it can take a minute for the algorithm to reduce the objective function from 200 000 to 5, but another half-hour to reduce it down to 4.95. This trend was repeated no matter what the optimisation scheme was used. Its performance pales in comparison to the GRG2 engine, which gives much faster results.

Substantial manipulation of the constraints had to be performed in order to obtain reasonable results. As the refinement proceeded, the mutation rate and the convergence tolerance were set to even more stringent values, to make sure that the algorithm sought out and considered as many varied potential solutions as possible. This resulted in a longer run time for the programme.

Another observed problem was that the algorithm tended to return extremely wild parameter sets, with ϕ and κ values in terms of 50 or 80. These values are clearly out of proportion, and the reason that the algorithm acts in such a way is that the constraints to the parameters are set to rather large ranges (E-7 to 100). Therefore, as the algorithm explores the feasible region it captures such ridiculous parameters. This is actually a sign that the algorithm is actively searching the entire feasible space, looking for alternative optima. This puts it at an advantage over the GRG2 algorithm, which may be trapped in a local optimum. The obvious solution to would be to change the constraints, but when these were reduced to more realistic values, there arose another problem. This time the algorithm tended to stick to the limits of the parameter constraints (especially for the parameters α and ν), again skewing the objective function.

Nevertheless, a few good results were obtained, and these can be seen in table 6.5 for November. Using the evolutionary algorithm does not appear to give better results than the normal deterministic methods. At most, it has obtained parameters that return an equally good objective function when compared to using the GRG2 engine. For the majority of the time, it returns inappropriate objective functions, for the reasons set above. It appears that the objective function is too complex for the genetic algorithm and that the GRG2 is a more efficient tool for optimisation. In the end, it was decided that the method was not suitable for use for the rest of the project, partly for its poor results and partly because the trial period for the downloaded software had come to an end.

Despite the poor results, the evolutionary algorithm cannot be totally disregarded in further studies. In this project, only a single package was used to perform the algorithm. Better results may be found by using different software packages or a different formulation of the objective function. In any case, genetic algorithms have had a mixed success rate its application, and since their development is still in its infancy, it should be used in conjunction with deterministic methods, rather than in place of them. Rapid developments in this field may lead to further breakthroughs, allowing new improved algorithms to be used, which may lead to better results.

7. Analysis of Extreme Data

7.1 Background and methodology

The Bartlett-Lewis Rectangular Pulse Model has been successful in the reproduction of a variety of statistics for a wide range of time scales, however one of its major deficiencies has been its inability to predict extreme values accurately (Onof and Wheater, 1993). Extreme statistics are needed for the estimation of return periods for extreme rainfall events. These values are used to fit extreme value distributions (Type I, Type II and Type III distributions) to find appropriate depths for the specified design life of the engineering project. For both the hourly and daily time scales, the model underestimated the extreme values, as it had not generated enough extreme rainfall events within the simulated period, as compared to historical data. This phenomenon is thought to relate to the fact that the model is not very accurate at reproducing the skewness of the time series. The model tends to return time series with a significantly lower skewness than the historical data, therefore indicating the lack of extreme values in the simulated series.

In section 6.3, it was noted for the modelled parameter sets that gave extremely accurate modelled hourly statistics, when these same parameters were used for disaggregation, the simulated time series returned an accurate skewness statistic. This is unusual, since no skewness or third order terms were used in the optimisation process. The accurate reproduction of the skewness after disaggregation indicated that extreme values might also be better reproduced by incorporating the disaggregation procedure.

The objective of the investigations presented in this section is to determine the extreme data characteristics of the disaggregated time series, and to find out whether they are better modelled after using a combination of the BLRPM and disaggregation.

Three months were chosen for analysis: March, August and December. These months were chosen for their spread throughout the year, incorporating both summer and winter months. Referring to fig. 6.8 in section 6.3, the modelled statistics for March are extremely well estimated, almost exactly the same as the historical data. The modelled statistics for December are reasonably estimated, while they are relatively poorly estimated for August. Therefore, the best results would be expected from March, followed by December and August.

Using the same parameter set as that used to derive the modelled statistics, Hyetos was used to simulate 39 years of monthly data, for each of the three months. For each month, this simulation was repeated thirteen times, each time using a different random seed. This generated thirteen different sets of simulated series based on the same parameter set, allowing a framework to be established to determine the spread of values caused by the choice of random seed. The thirteen sets of data are plotted on a logarithm plot so that the spread (maximum and minimum) of values can be easily seen, and this is compared to the actual historical data.

For each disaggregation, the time series output was entered into a spreadsheet. For each year, the maximum rainfall depths for the 1-hr, 6-hr and 12-hr time scales were determined. Using this 39 year maximum value data, the FORTRAN programme `gev` was used to fit an extreme value distribution to the maximum value data. There are three possible distributions available, the Type I (Gumbel) distribution, the Type II or the Type III plot. The goodness of fit is given by a K-value, where a value close to zero indicates a good fit to the Type I distribution, a K-value less than zero indicates a Type II distribution, and a positive K-value denotes a Type III distribution. The programme gives an output of the rainfall depths with respect to the Gumbel reduced variate, which is related to the return period according to the equation:

$$\text{Gumbel Reduced Variate (GRV)} = -\ln(-\ln(1-1/T)) \qquad (9)$$

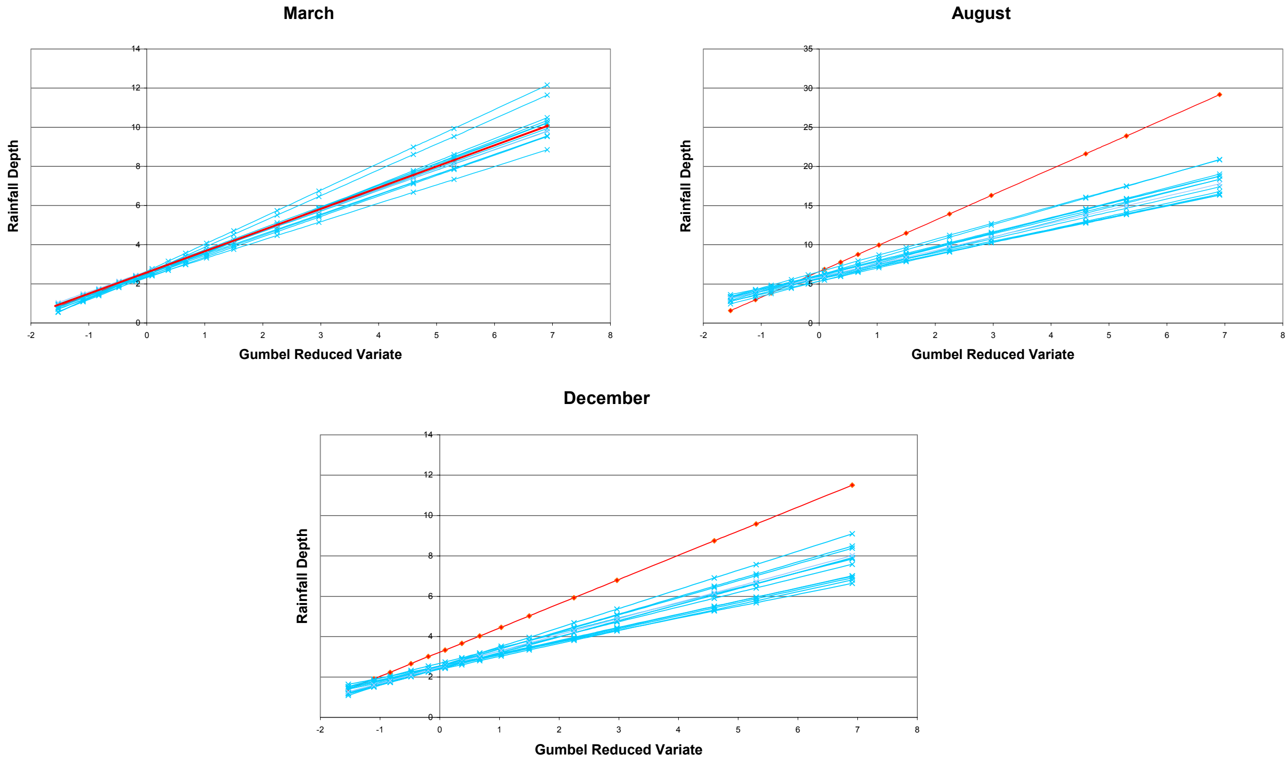
where T is the return period. This entire process is then repeated for the 6-hr and 12-hr time scales for each month. The actual historical data are also processed to obtain similar plots, and these are used to compare with the simulated plots.

7.2 Analysis results

Extreme values were determined for 1-hr, 6-hr and 12-hr intervals for each of the three months. A Type I (Gumbel) distribution was used to plot all the data points, since the majority of the K-values fell between -0.2 and 0.2. The plots show the rainfall depth against the Gumbel reduced variate, which is related to the return period of the rainfall depth.

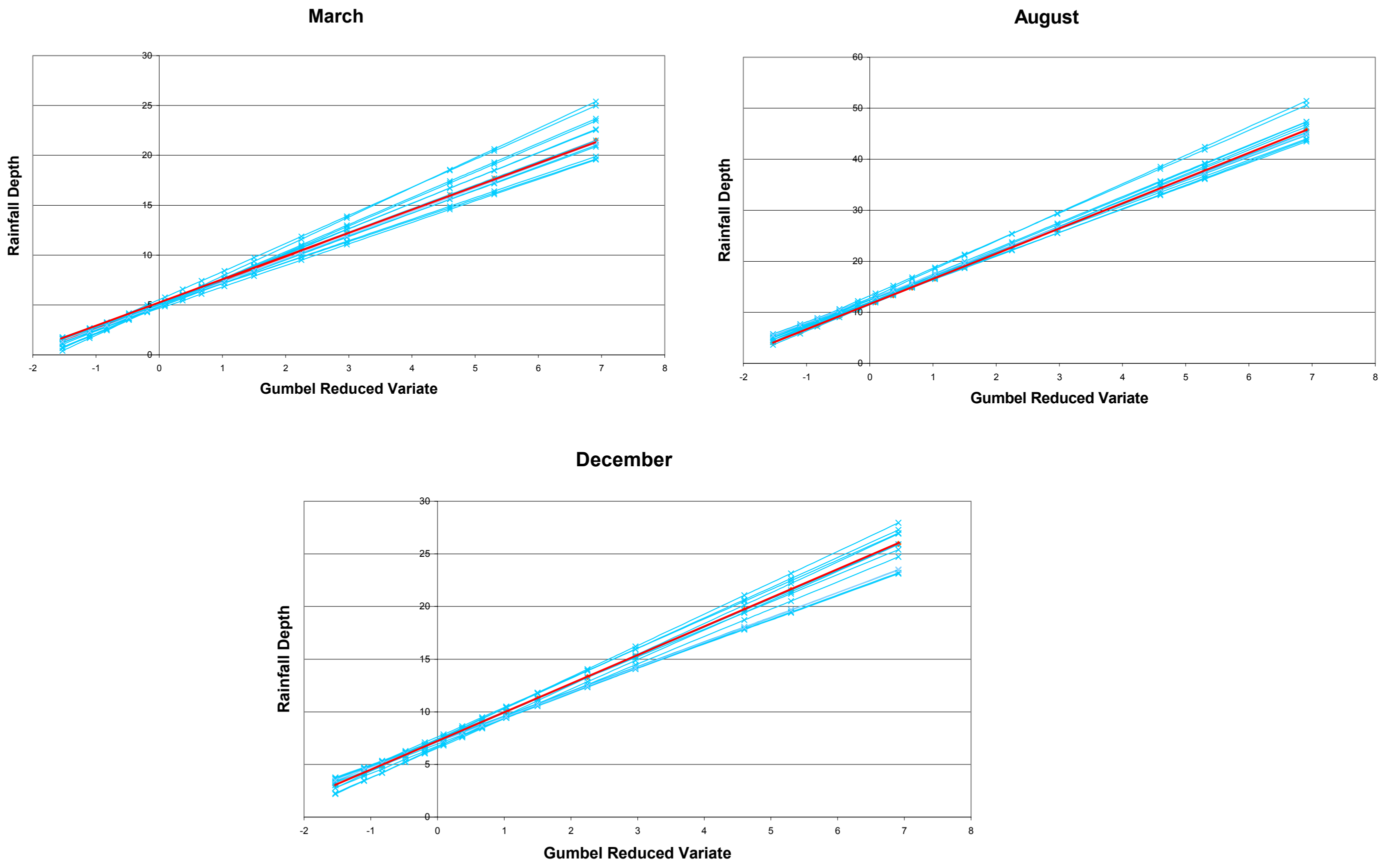
Fig. 7.1 shows the hourly extreme values for the three months considered. The blue lines show the thirteen simulations performed, and the maximum and minimum of the spread can be estimated from these plots. The plots diverge steadily as the return period increases. The red line shows the plot for the historical statistics. The data and K-values for the simulation can be found in Appendix E.

Figure 7.1: 1-hr extreme values



As expected, the results for March are the best, with the historical values lying directly in the middle of the spread of simulated values. This is very encouraging, since it shows that if the parameters derived from pure daily data can model the hourly statistics accurately, it can also give very accurate estimates for the extreme values of rainfall. For August and December, the parameters did not model the historical statistics as accurately, and this is reflected in the plots for the extreme values. The historical values show a larger number of extreme rainfall events than the simulated series. The result could be due to the inaccurate BLRPM parameters being used or may simply indicate that the BLRPM is unable to reproduce enough extreme events. The close relation between the skewness and the extreme values would indicate that corresponding results would be obtained for the skewness statistic.

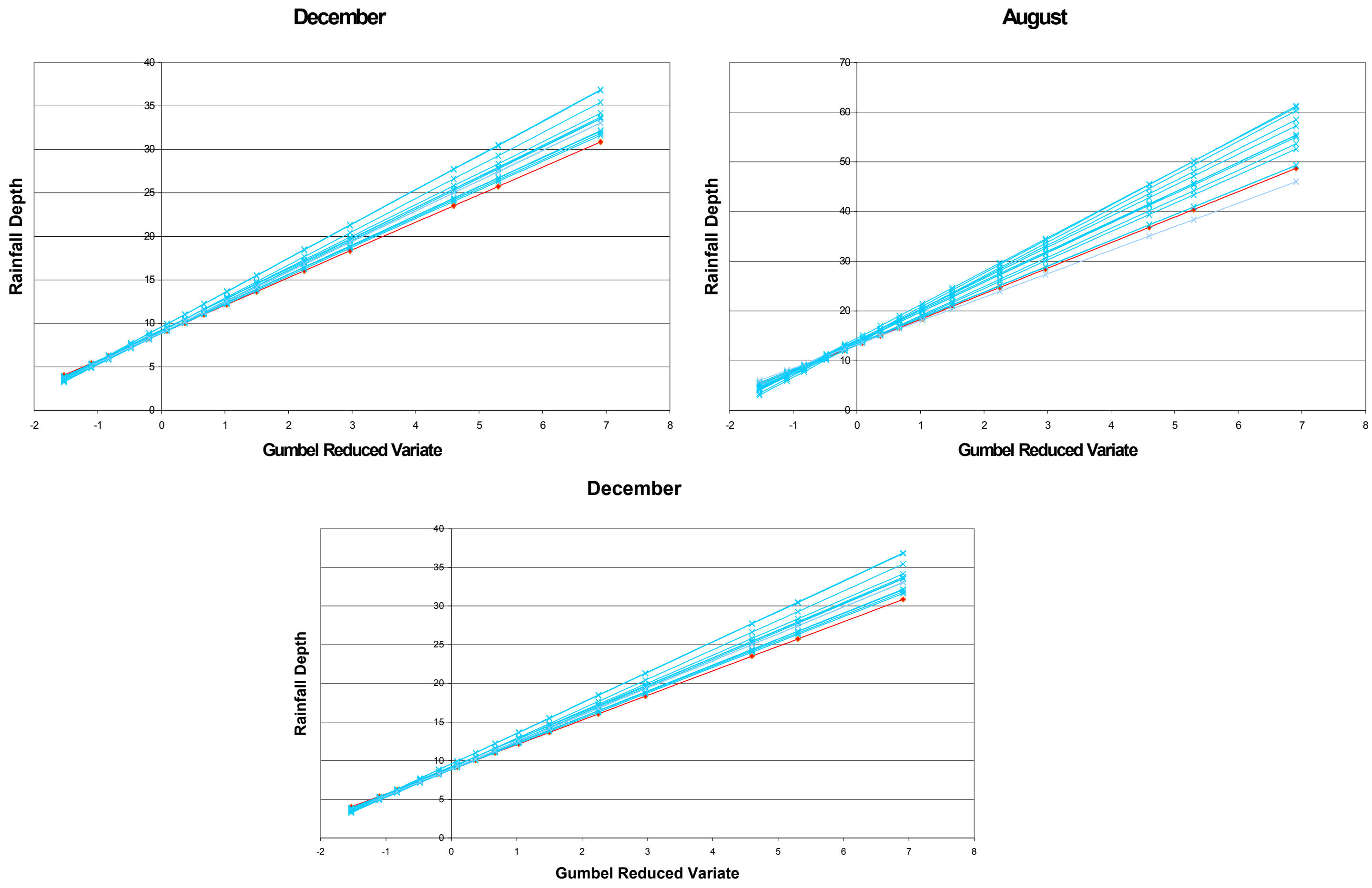
Figure 7.2: 6-hr extreme values



The extreme 6-hr extreme values are more surprising, as show by fig. 7.2. The results show that the 6-hr extreme values are very accurately estimated for every month. The historical plots lie straight in the centre of the spread of simulated series. This appears to indicate that the disaggregation procedure has actually improved the estimation of 6-hr extreme values.

The 12-hr extreme value results are also encouraging, as shown in fig. 7.3. For this time scale, the historical plot tends to lie at the lower end of the spread of the simulated series. This is in reverse to all the previous observations, since the simulations have now generated more extreme rainfall events than required! Hence, simulation can give a reasonable, albeit slightly higher estimation of the extreme value plot.

Figure 7.3: 12-hr extreme values



The Type 1 distribution is a good representation of the distribution of the extreme values. The majority of the K-values for the simulations lay within the range of -0.2 and +0.2, indicating a Gumbel distribution. The simulations that had K-values beyond this range tended to fall below the stipulated range, (but never beyond -0.4), indicating that for these particular simulations, a Type 2 distribution may be more appropriate. Nevertheless, these values were plotted using a Type 1 distribution for consistency.

In summary, using disaggregation in combination with the BLRPM appears to give better estimates for the extreme values than using the BLRPM alone. This result is in agreement with the paper by Onof and Koutsoyiannis (2000). This is most significantly reflected in the 6-hr data, which are excellently reproduced. The 12-hr data are also reasonably estimated, although slightly high. The 1-hr extreme values give a variety of results. It appears to be the trend that if the modelled statistics are well estimated, the extreme values will follow in suite. Otherwise, the 1-hr statistics will tend to be underestimated, with insufficient extreme hourly rainfall events being generated.

The reason behind this improvement can be attributed to the repetition stage of the disaggregation process. As described in section 5, the disaggregation process involves repetition of the simulation, so that better high level statistics are obtained. Performing this repetition makes the chances of obtaining good estimates for the statistics that are not explicitly preserved much higher. Skewness, and therefore the extreme value statistics, is improved from the original Bartlett-Lewis rainfall model.

The process has yet to be repeated for the rest of the months to check for any anomalies. Since only a simple proportional disaggregation procedure is used, better results may be found by using a different disaggregation procedure, such as the linear or power adjustments. Other disaggregation procedures may be coupled with the BLRPM to improve the hourly extreme values, and further research is required in this area.

8. Power Spectrum Analysis

8.1 Introduction and Methodology

The power spectral density function, or power spectrum for short, shows the contributions of various frequencies to a process. Therefore, for a given rainfall time series, the power spectrum will show the frequency decomposition of the variance of the series, showing the relative effect each frequency has on variability of the process. Formally, the power spectrum is the Fourier transform of the autocovariance process:

$$f^{T}(\omega)=\frac{1}{\pi}\int_{\infty}^{\infty}\gamma^{T}(s)\; e^{-i\omega s}\,ds \qquad 0<\omega<\infty \qquad (10)$$

where T is the time scale being considered, and γ is the autocovariance function.

Power spectrum analysis has been carried out on rainfall time series by Bo, Islam and Eltahir (1994), working with rainfall data from Italy and Kentucky, U.S.A. Using the same rainfall model together with disaggregation, they found that the power spectrum showed two distinct regions, firstly a flat section among the low frequencies, and then a negative sloping curve as the frequencies increased. The corner frequency that separated the two regions was found to be related to the rate of storm arrival, or λ. The flat region in the power spectrum was of particular interest, since this indicated that below the corner frequency, the contribution from each frequency would be roughly equal. Therefore for such low frequencies, the spectrum is virtually equivalent to a spectrum for white noise. Working with both disaggregated and aggregated data, this research group managed to find that the power spectrum closely matched each other. Hence they inferred that the BLRPM is self-consistent, in that despite using information from a high level time scale, (though not beyond the corner frequency), the model could estimate to high accuracy statistics at lower time scales.

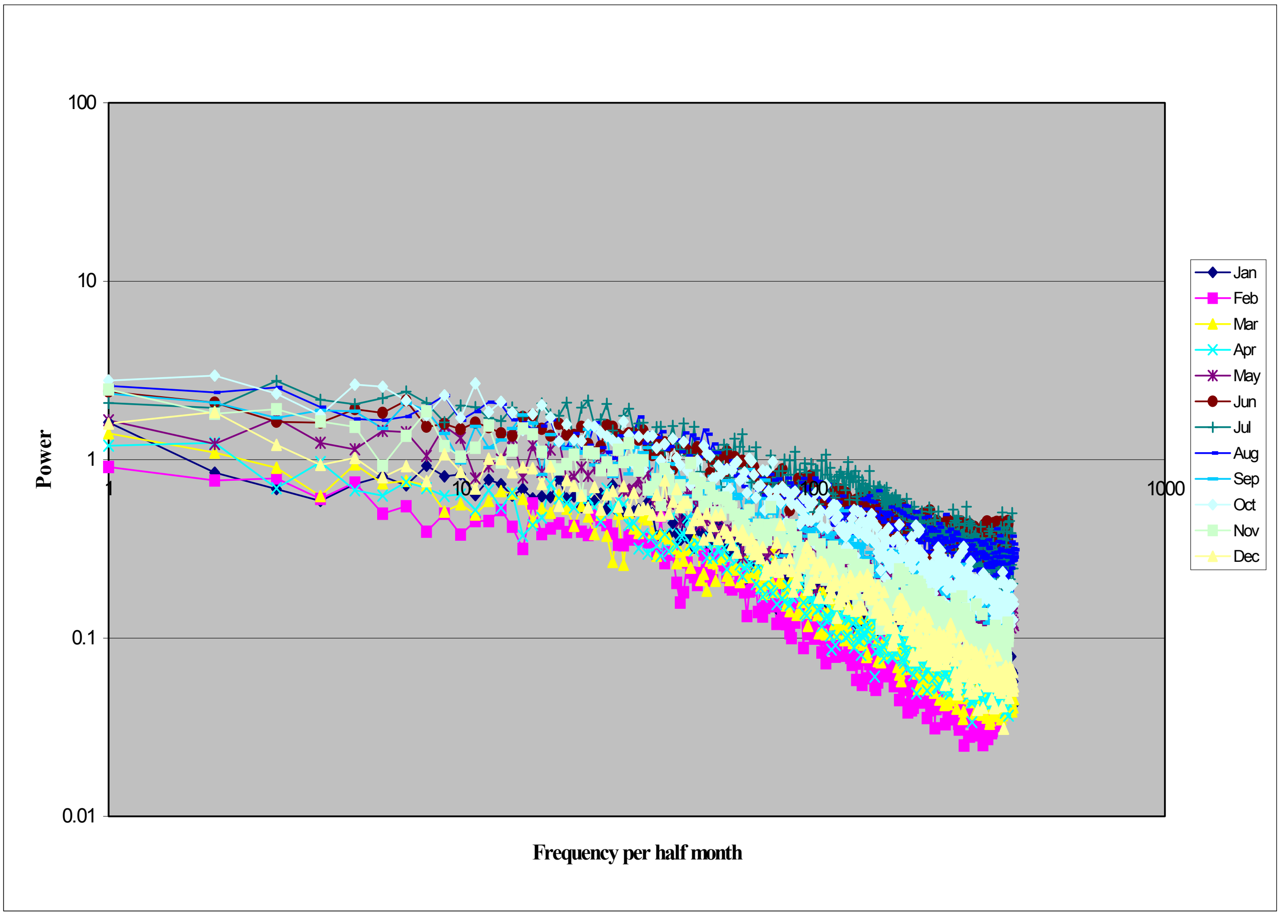
These findings seem very much in agreement to the findings set forth in this paper, in that the use of coarse time scale statistics can be used to estimate accurate fine time scale statistics. Further evidence for these empirical findings can be offered by analysis of the power spectrum. If the power spectrum of the historical data is close to the simulated plot, then this will re-emphasise the observations presented before.

A numerical method is needed to evaluate the power spectrum. This was done using the FORTRAN programmes `pg/mdtomo/gaugebin`. The programme uses the time series entered to calculate the Fourier transform, and gives the modulus of the real and imaginary terms, with respect to the frequency. The result is a plot of power against frequency, the frequency being the number of times every half a month (372 hours).

8.2 Analysis of Results

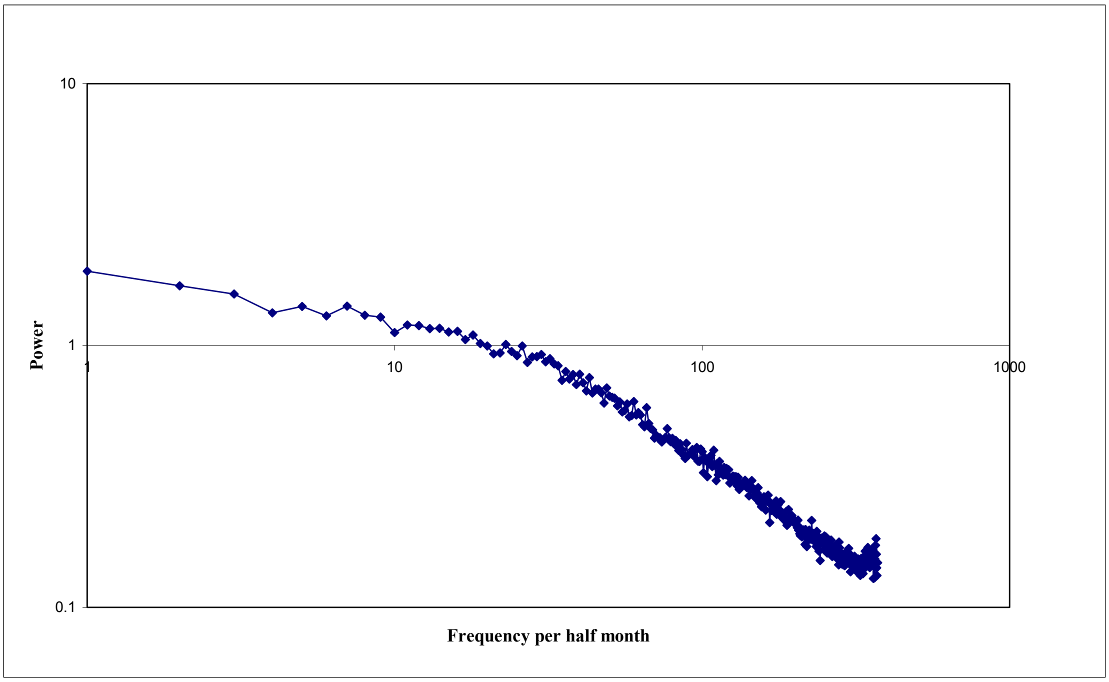
A computer programme, `pg/mdtomo/gaugebin` was used to derive the power spectrum for each and every month for the historical time series. The results are plotted in fig. 8.1. Note that both axes are logarithmic, and that the x-axis is in terms of frequency per half month.

Figure 8.1: Monthly power spectrum for historical data



The power spectrum shows similar characteristics to those presented in Bo, et al. (1994). There is a flat region corresponding to the random "white noise" spectrum for lower frequencies, and a negative sloping region for high frequencies. This is perhaps more evident in fig. 8.2, which shows a simple monthly-averaged power spectrum. However, it is noted that the individual months' spectra show a much flatter initial portion; the averaged spectrum appears to still have a slightly downward sloping curve in its initial portion. This effect may be due to the averaging of the spectra over the months, including the fact that different months contain different numbers of days, which may lead to distortions in the curve. It is advisable to use individual months' spectra for comparison with simulated data rather than the averaged spectrum, although this latter spectrum is useful for general observations.

Figure 8.2: Averaged spectrum for historical data



Considering fig. 8.2, the flat portion of the spectrum consists of "random white noise" and therefore it may be concluded that such low frequencies do not help the process of estimating low level statistics. The corner frequency is important, and this occurs between 8 and 30 in the figure above, corresponding to periods of 46.5hrs and 12.4 hrs. There exists a range where the corner frequency can occur due to the distortion effect of smoothing of the twelve monthly spectra. Bo et al. (1994) related the corner frequency to the storm arrival rate, λ, and this finding appears to be proven true in the Heathrow data as well. The range of corner frequencies in the figure above implies λ values between 0.08065 and 0.0215, where the corner frequency is the inverse of the storm arrival rate. The majority of the λ values derived under the various optimisation schemes do indeed fall within this range. However, most of the λ values cluster around the lower end (≈ 0.02) of this range (see the BLRPM parameter charts at the bottom of the data tables in Appendix C), indicating that the corner frequency may be lower than the graph indicates. Observations from fig. 8.1 indicate that most of the individual monthly spectra have corner frequencies around 6 - 9 per half month. These values correspond to a range of λ between 0.01613 and 0.02494, and are nearer to the observed λ values from the optimisations.

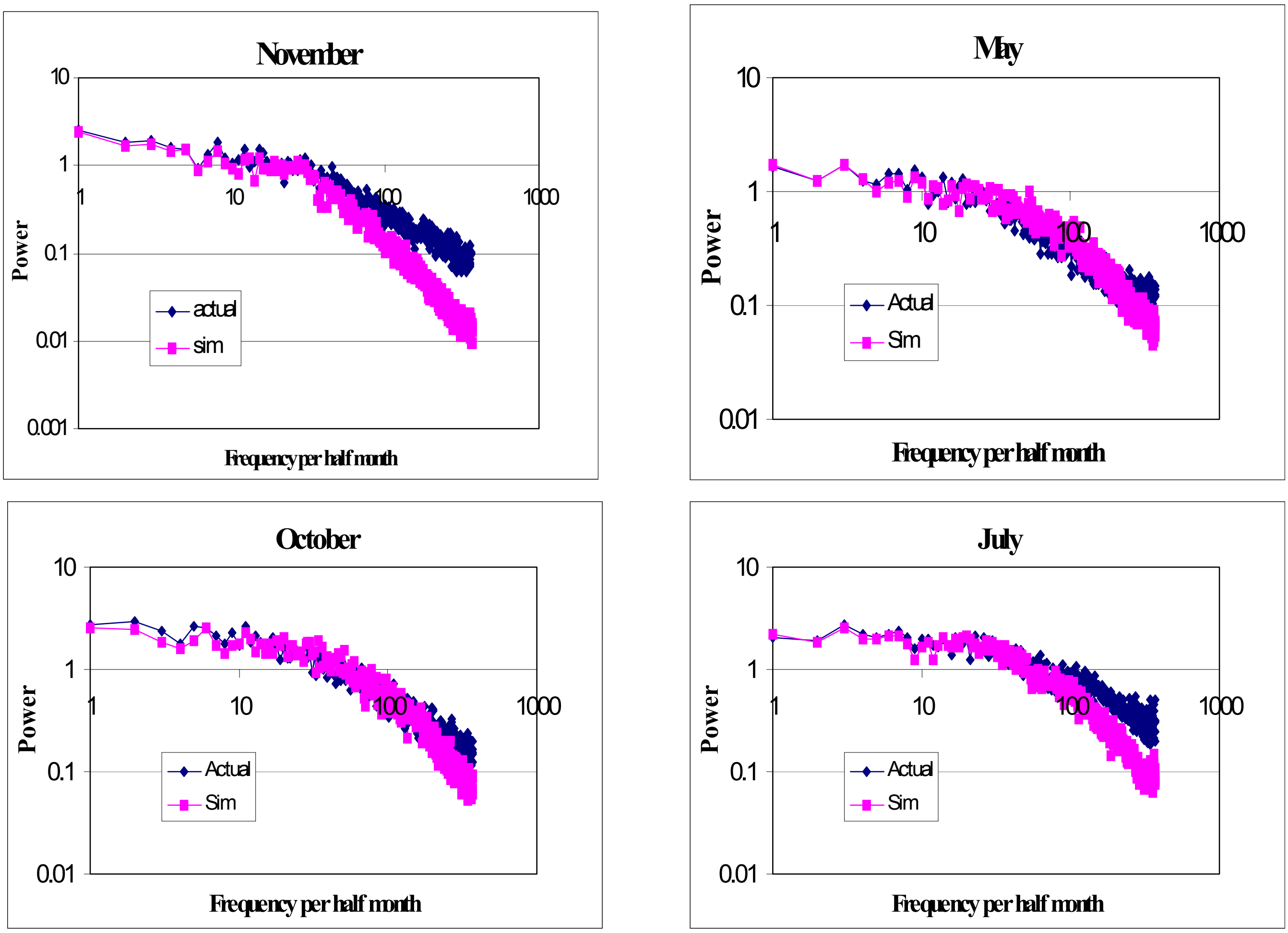
The observations indicate that the frequencies below the corner frequency are not useful for the estimation of results, since they correspond to what is essentially random noise. Only statistics from the frequencies above the corner frequency, that is, periods below 48 hours, are suitable.

It is interesting to note the spread of the spectra between the various months in fig. 8.1. The spectra corresponding to winter months tend to occur lower than those corresponding to the summer months. There is a clear upward shift in the spectra from winter to summer. This seems to imply the variability arising from all the frequencies increases during the summer. This might have some connection to the variable results obtained for the summer months during the optimisation, especially the underestimated variance statistics.

The programme was then modified to accept simulated time series. Time series were generated using disaggregation (Hyetos), for each month. The parameters were the same as those used to derive the modelled statistics in fig. 6.8. The power spectra for these series were then used to compare with the historical spectra.

The important test for the disaggregation procedure is whether the simulated disaggregated time series can reproduce the power spectrum observed in the historical data. Time series generated by Hyetos were used as input into the programme to find the simulated time series' power spectrum. This was done for several months and the simulated spectra were compared to the historical data. Fig. 8.3 shows plots for various months.

Figure 8.3: Simulated and historical spectra



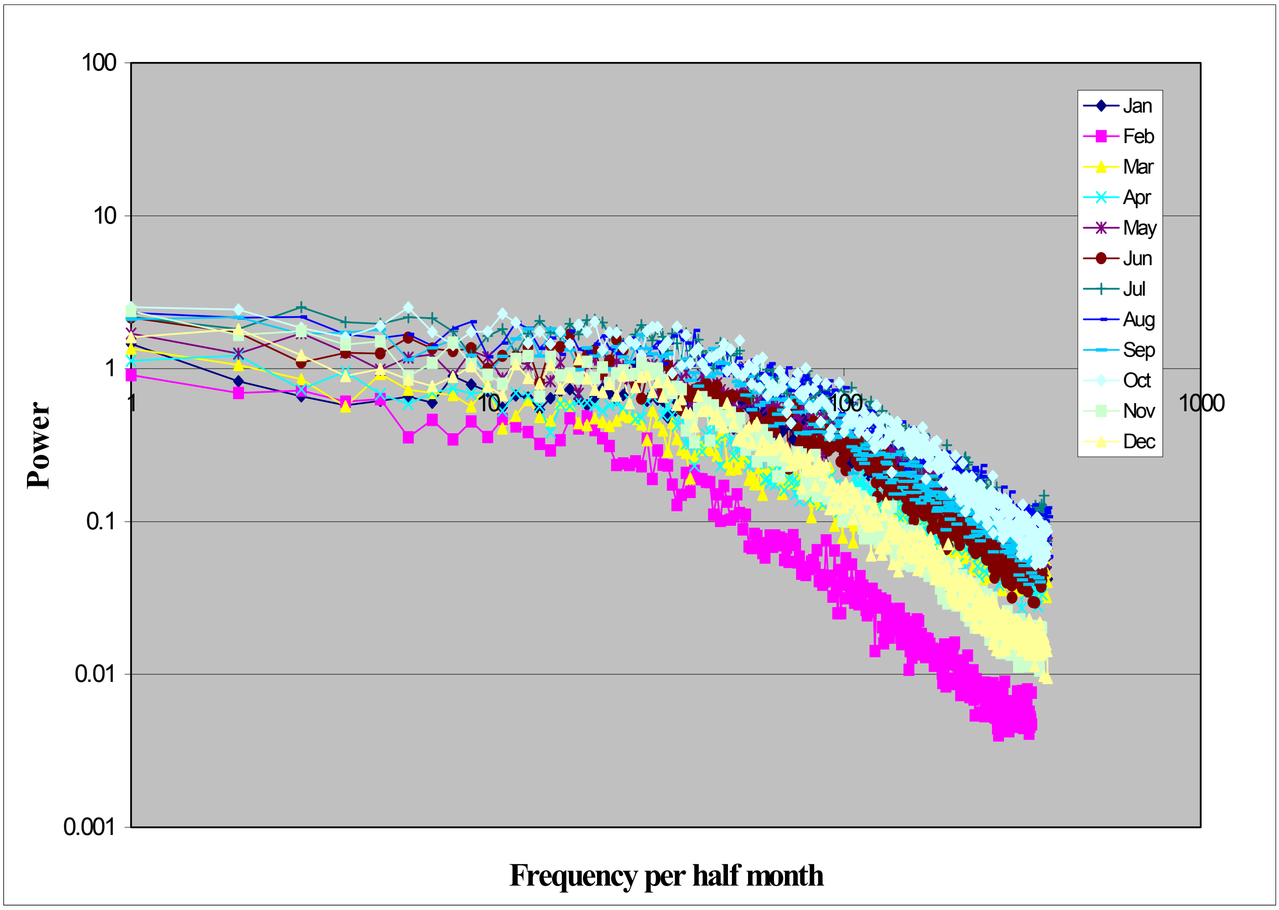
It is clear from these plots that the simulated spectra do indeed following closely to the historical spectra. The effect is most significant in the lower frequencies, and this is due to the fact that these frequencies correspond to periods of days and higher. The statistics are completely preserved for such time scales since they are higher than the historical time scales available. No disaggregation is needed to obtain such statistics, the statistics can be derived exactly by aggregation. For several months, the two spectra are almost exact.

Divergence does occur for other months, and this is rather worrying, since it occurs at the high frequencies corresponding to periods of about 1 hour, exactly the time scale that is being estimated. Divergence usually occurs around a frequency value of about 120 per half month, corresponding to a period of roughly 3 hours. This is quite near the hourly time scale being aimed for, therefore disaggregation should reasonably preserve the characteristics of the series down to this time scale.

From this observation, it could be concluded that, using only coarse time scale data, statistics from the intermediate time scales of 6-hr and 12-hr can be estimated very accurately. Modelled statistics for such time scales are shown in Appendix B, and this conclusion does hold true to some extent. It is observed that for statistics that are estimated with low accuracy at the hourly time scale, the 6-hr and 12-hr estimates are increasingly better estimated. Evidence from the extreme values is also encouraging, since 6-hr and 12-hr extreme values are well estimated, while 1-hr values being more variable. This could be related to the divergence of the power spectra occurring at the 3 hour time scale for the various months, resulting in good values for the 6-hr and 12-hr statistics where there is convergence, and poor results for the hourly statistics where there is a slight divergence. There are of course some anomalies to this case, and further investigation must be done in this area.

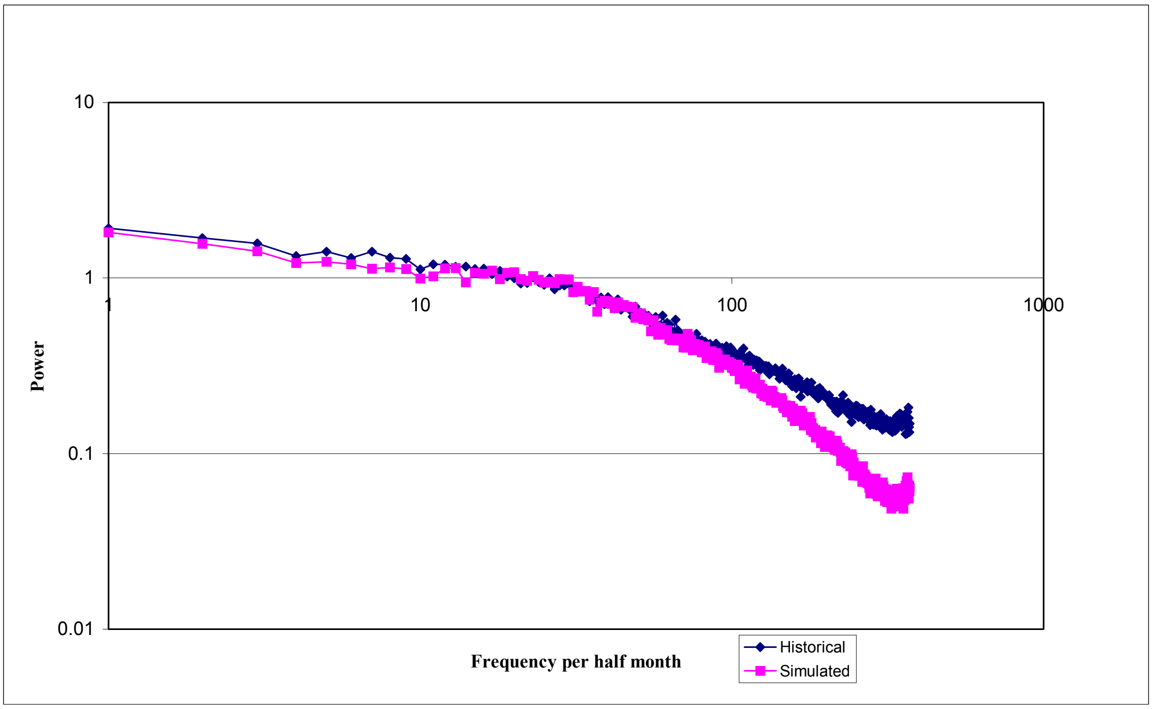
The monthly simulated power spectra for the entire year are shown in fig. 8.4 below.

Figure 8.4: Monthly simulated power spectra



The power spectra derived from the simulated time series have the same shape as the actual spectra. There is also some evidence of the summer months' spectra being located in the upper portions of the range, while the winter months gathering at the lower end of the range of values. This mirrors the observations made before in the historical spectra. When the simulated months are averages out, the result is a plot shown in fig. 8.5.

Figure 8.5: Averaged simulated and historical spectra



A clear divergence is shown between the two spectra. The effect is somewhat exaggerated, since in actual fact, only four months showed significant divergence. Two of the remaining months were reasonable well correlated, while for the rest of the half year, the simulated and actual power spectra were also identical. Of particular significance was the spectra for June, which had quite a large divergence. This month had previously shown the least accurate estimated statistics, and therefore the divergence is no surprise. The result simply reconfirms the fact that diaggregation has not been able to adequately reproduce a consistent rainfall structure for this month.

It must also be noted that divergence inevitably occurred for those months for which some error appearing during the computer programme, such as the programme not being able to model a particular storm with special characteristics. Such errors may affect the final results, in turn passing on the error to the power spectra. Therefore divergence may be a result of poor disaggregation. Using a range of random seed for the simulation programme also appears to give only minor changes to the power spectra.

The close relation between the simulated and actual spectra shows that the disaggregation procedure is consistent at reproducing the underlying variability structure behind the time series. This gives further merit to the use of disaggregation, since if this underlying structure may be found, the resulting simulated statistics are likely to be well estimated. It would be difficult to try and model the power spectrum, since its analytical form is rather complicated (Bo et al., 1994) therefore it would probably be impractical to model a power spectrum and work backwards to find the statistics required.

9. Conclusions

Hourly rainfall data is often required for engineering or hydrological purposes, but is also often severely lacking, both in terms of spatial coverage as well as length of recorded time. Daily rainfall data is readily available, however, and this paper examines a method of obtaining such hourly time series from daily data. The randomised Bartlett-Lewis Rectangular Pulse Model, in combination with disaggregation using adjusting procedures, appears to be able to estimate hourly statistics to a reasonable accuracy. The main conclusions can be summarised in the following points:

1. The six parameter Bartlett-Lewis Rectangular Pulse Model is used together with a proportional adjusting procedure in the disaggregation process. These are used to simulate a low level time series in order to compare with the actual hourly data available. Data from an hourly raingauge at Heathrow Airport, U.K. was used.
2. Parameter fitting of the Bartlett-Lewis Rectangular Pulse Model requires the use of various historical statistics from the daily or 48-hr (coarse) time scales. Optimisations were performed using 4, 6 and 8 statistics. The optimisations using four or six statistics showed the best reproduction of hourly (fine) time scale statistics. Using eight statistics incurred significant error, which may be attributed to a larger number of statistics being used to estimate fewer parameters.
3. Using six statistics gives better results than using four statistics. The use of six statistics is the most theoretically sound, since there are an equal number of statistics and parameters being estimated. However, the use of only four daily statistics is only slightly less accurate, and has the advantage of being faster and being purely dependant on 24-hr statistics.
4. The optimisation procedure is strongly affected by the initial points selected, as the feasible region is complex and contains local minima. A large range of initial points must be tried to identify the local minima present. Often the majority of the initial points would converge to a local minimum that returned the lowest objective function value among all the trials performed. This was termed the lowest local minimum.

5. The parameters obtained from this lowest local minimum showed distinct characteristics. The modelled statistics derived from these parameters were compared to the historical data and it was found that the mean, autocovariance and proportion of dry periods were well estimated. The variance was only reasonably estimated for about half the months. In particular, the summer months showed a strong underestimation of the historical variance. It is thought that this might relate to the relatively higher power spectra shown by the summer months.

6. Other parameter sets that did not belong to the lowest local optimum (i.e. having a higher objective function) were also used to derive modelled statistics. These statistics were on the whole poor estimates as compared to the ones obtained previously, and did not accurately model the historical data. There was, however a significant minority of modelled statistics that were highly accurate. This latter observation is probably due to the fact that there is a disparity between the objective function using an optimisation with pure hourly data, and an optimisation with pure daily data. Therefore, the highly accurate statistics may be due to the optimisation finding a local minimum that may be lowest under optimisation with pure hourly data, but not with pure daily data.

7. The importance of the initial values is re-emphasised by the investigations in Case 3, where the initial values are modified by the hourly data prior to the optimisation using only daily data. This case produced the most accurate results for all the months and all the statistics. Although highly accurate, this method cannot be used in practise, since it still requires some hourly input. It may be possible to find some correlation within a network of rain gauges, that incorporates some hourly rain gauges, between the gauges with hourly readings and those without. Hourly statistics from a neighbouring rain gauge may be used to modify the initial values for an optimisation using statistics from a daily rain gauge. This is an area that requires further research.

8. The use of disaggregation was vital in producing the actual simulated time series. Disaggregation itself seemed to give only a small improvement on the modelled statistics, as the simulated statistics followed the modelled statistics closely.

9. For parameter sets that gave very accurate modelled statistics, when these were used in disaggregation, the skewness was particularly well estimated. This is

encouraging since the skewness statistics has often been underestimated in the BLRPM.

10. The evolutionary algorithm may be used to overcome the problem of finding a suitable initial value. However, the engine used in this study proved too slow and ponderous, producing poor results at an inefficient rate. This may indicate the need for a reformulated objective function to better suit the needs of the algorithm.

11. The extreme values for the simulated time series were particularly well estimated for the 6-hr and 12-hr time scales. This shows that the repetition process in the disaggregation procedure has a positive effect in improving the lack of extreme rainfall events that was a deficiency in the BLRPM. Varied results are returned for the 1-hr extreme events. It appears that the 1-hr extreme data is only well represented if the other statistics are modelled with a high accuracy.

12. The power spectrum of the time series consists of two parts: a flat portion in the lower frequencies that appears to be random white noise, and a negative sloping portion in the higher frequencies that contains the variability structure of the series.

13. The simulated power spectrum was well reproduced by disaggregation and lay close to the actual spectrum. This indicates that the disaggregation procedure preserves the underlying variability structure of the rainfall time series. Any divergence occurs at a frequency corresponding to a period of three hours, therefore indicating that coarse statistics may be used to estimate accurately statistics up to the three hour time scale. This time scale is close the hourly target. The fact that the divergence is small indicates that hourly statistics may still be reasonably well estimated using coarse statistics.

In conclusion, the coupling of the randomised Bartlett-Lewis Rectangular Pulse Model and the proportional adjusting disaggregation procedure proves to be a good framework in which hourly data can be simulated from daily statistics. Apart form some deficiencies in the hourly variance and extreme value plot, the mean, autocovariance and proportion of dry periods are often well estimated. These results seem to be re-emphasised and justified by the close relation between the simulated and actual power spectra. The use of the Bartlett-Lewis rainfall model in conjunction with disaggregation appears to be very promising for the derivation of hourly rainfall time series from daily statistics.

Suggested areas of further research:

1. Case 3 shows a strong case for the modification of the initial values by hourly statistics, before the optimisation is performed. In practise, this case is impossible to perform since the hourly statistics are precisely the statistics that are being estimated. It may be possible to find some correlation within a network of rain gauges, that incorporate some hourly rain gauges, between the gauges with hourly readings and those without. Therefore, hourly statistics from a neighbouring rain gauge may be used to modify the initial values for an optimisation using statistics from a daily rain gauge.
2. The feasible region is highly complex and a different objective function may be used to elucidate its nature and the location of the local minima. A suggested objective function would consist of the error terms being expressed in percentage terms, instead of the absolute values of error used in this paper. A similar study could be carried out on the optimisation process by varying the weights applied to each statistic.
3. The use of the evolutionary algorithm in this paper was very basic and a different type of algorithm (such as simulated annealing or neural networks) could be used in another study. The feasible region and the objective function could be tailored to suit the algorithm, or an entire algorithm customised for the Bartlett-Lewis rainfall model, could be written.
4. Only three months of extreme value data was considered, due to time constraints. The remaining months may be analysed, and other simulated time series from different optimisations could be examined as well. The links between the extreme values and skewness can be elucidated.
5. More work needs to be performed on the analysis of power spectra developed from disaggregated simulations. Special attention should be applied to 6-hr and 12-hr modelled statistics.

10. References

1. Bo, Z., S. Islam and E. A. B. Eltahir, (1994) Aggregation-disaggregation properties of a stochastic rainfall model, *Water Resources Research, 30(12)*, 3423-3435.
2. Cryer, J. D. (1986) Time series analysis, Massachusetts: PWS Publishers.
3. Frontline Systems, Inc. (1999) http://www.frontsys.com
4. Koutsoyiannis, D., and A. Manetas, (1996) Simple disaggregation by accurate adjusting procedures, *Water Resources Research, 32(7)*, 2105-2117.
5. Koutsoyiannis, D. and C. Onof, A computer program for temporal rainfall disaggregation using adjusting procedures, XXV General Assembly of European Geophysical Society, Nice, *Geophys. Res. Abstracts*, ?(?), p. ?, 2000. (Presentation also available on line at http://www.hydro.ntua.gr/faculty/dk/ pub/Hyetos.pdf)
6. Koutsoyiannis, D. and C. Onof (2000) Rainfall Disaggregation using adjusted procedures on a Poisson cluster model, *Pending publication.*
7. Lasdon, L. S. and A. D. Waren (1998) Optimal Methods Inc. http://www.optimalmethods.com.
8. Onof, C., and H. S. Wheater, (1993) Modelling of British rainfall using a random parameter Bartlett-Lewis Rectangular Pulse Model, *Journal of Hydrology, 149*, 67-95.
9. Onof, C. and H. S. Wheater, (1994) Improvements to the modelling of British rainfall using a modified random parameter Bartlett-Lewis Rectangular Pulse Model, *Journal of Hydrology, 157*, 177-195.
10. Rodriguez-Iturbe, I., D. R. Cox, and V. Isham, (1988) A point process model for rainfall: Further developments, *Proceedings of the Royal Society of London A, 417*, 283-298.
11. Rodriguez-Iturbe, I., D. R. Cox, and V. Isham, (1987) Some models for rainfall based on stochastic point processes, *Proceedings of the Royal Society of London A, 410*, 269-288.

Acknowledgements

I would first like to thank Pamela Wright from Frontline Systems, Inc., who was most helpful in providing the software for the evolutionary algorithm, and Dr. Christian Onof who has been so patient and resourceful throughout these few months. Last but certainly not least, I would also like to thank Professor Demetris Koutsoyiannis whose inspirational emails and insightful comments helped immensely when the going was tough.

11. Bibliography

1. Bo, Z., S. Islam and E. A. B. Eltahir (1994) Aggregation-disaggregation properties of a stochastic rainfall model, *Water Resources Research, 30(12)*, 3423-3435.
2. Bras, R. L. and I. Rodriguez-Iturbe (1993) Random functions and hyrdrology, 2nd Ed., New York: Dover Publications.
3. Chatfield, C. (1989) The analysis of time series, 4th Edition, Great Britain: Chapman and Hall Ltd.
4. Cryer, J. D. (1986) Time series analysis, Massachusetts: PWS Publishers.
5. Davis, L. (1991) Handbook of genetic algorithms, New York: Van Nostrand Reinhold.
6. Frontline Systems, Inc. (1999) http://www.frontsys.com
7. Heitkoetter, J. and D. Beasley, (1994) The Hitch- Hiker's Guide to Evolutionary Computation: A list of Frequently Asked Questions (FAQ), USENET: comp.ai.genetic. FTP: rtfm.mit.edu:/pub/usenet/news.answers/ai-faq/genetic/.
8. Hillier, F. S. and G. J. Lieberman (1995) Introduction to operations research, International Edition, Singapore: McGraw-Hill Book Co.
9. Jacobsen, R. (1997) Microsoft Excel 97: Visual Basic Step by step, Redmond, Washington: Microsoft Press.
10. Koutsoyiannis, D. and A. Manetas, (1996) Simple disaggregation by accurate adjusting procedures, *Water Resources Research, 32(7)*, 2105-2117.
11. Koutsoyiannis, D. and C. Onof, A computer program for temporal rainfall disaggregation using adjusting procedures, XXV General Assembly of European Geophysical Society, Nice, *Geophys. Res. Abstracts*, ?(?), p. ?, 2000. (Presentation also available on line at http://www.hydro.ntua.gr/faculty/dk/ pub/Hyetos.pdf)
12. Koutsoyiannis, D. and C. Onof (2000) Rainfall Disaggregation using adjusted procedures on a Poisson cluster model, *Pending publication.*
13. Lasdon, L. S. and A. D. Waren (1998) Optimal Methods Inc. http://www.optimalmethods.com.
14. Montgomery, D. C. (1999) Applied statistics and probability for engineers, 2nd Ed., Chichester: Wiley.
15. Onof, C., and H. S. Wheater, (1993) Modelling of British rainfall using a random parameter Bartlett-Lewis Rectangular Pulse Model, *Journal of Hydrology, 149*, 67-95.
16. Onof, C. and H. S. Wheater, (1994) Improvements to the modelling of British rainfall using a modified random parameter Bartlett-Lewis Rectangular Pulse Model, *Journal of Hydrology, 157*, 177-195.
17. Priestley, M. B. (1992) Spectral analysis and time series, 7th Edition, Great Britain: Academic Press Ltd.
18. Quagliarella, D. et al., (1998) Genetic algorithms and evolution strategy in engineering and computer science: recent advances and industrial applications, Chichester: Wiley.
19. Rodriguez-Iturbe, I., D. R. Cox, and V. Isham, (1988) A point process model for rainfall: Further developments, *Proceedings of the Royal Society of London A, 417*, 283-298.
20. Rodriguez-Iturbe, I., D. R. Cox, and V. Isham, (1987) Some models for rainfall based on stochastic point processes, *Proceedings of the Royal Society of London A, 410*, 269-288.
21. Wall, M. (2000) GAlib: Mattew's Genetic Algorithm Library, http://lancet.mit.edu, FTP: lancet.mit.edu/pub/ga.
22. Webb, J. (1996) Using Excel Visual Basic for Applications, Indianapolis: Que Corp.
23. Wheater, H. S., V. S. Isham, D. R. Cox, R. E. Chandler, A. Kakou, P. J. Northrop, L. Oh, C. Onof, and I. Rodriguez-Iturbe, (1997) Spatial-temporal rainfall fields: modelling and statistical aspects, http://www.ucl.ac.uk/Stats/research/rr176/176.html.

APPENDIX A

COMPUTER PROGRAMS

OPTIMA

`Optima` was developed for the needs of the project and its main features and algorithm is explained here. It uses a SOLVER engine to perform the optimisations, which can be considered as a black box. Because of the nature of Visual Basic within Excel, it is rather difficult to reproduce the full code in text form, as the coding is placed in many different parts of the programme. Two sister programs, `norway` and `evo` were also produced in the development of optima.

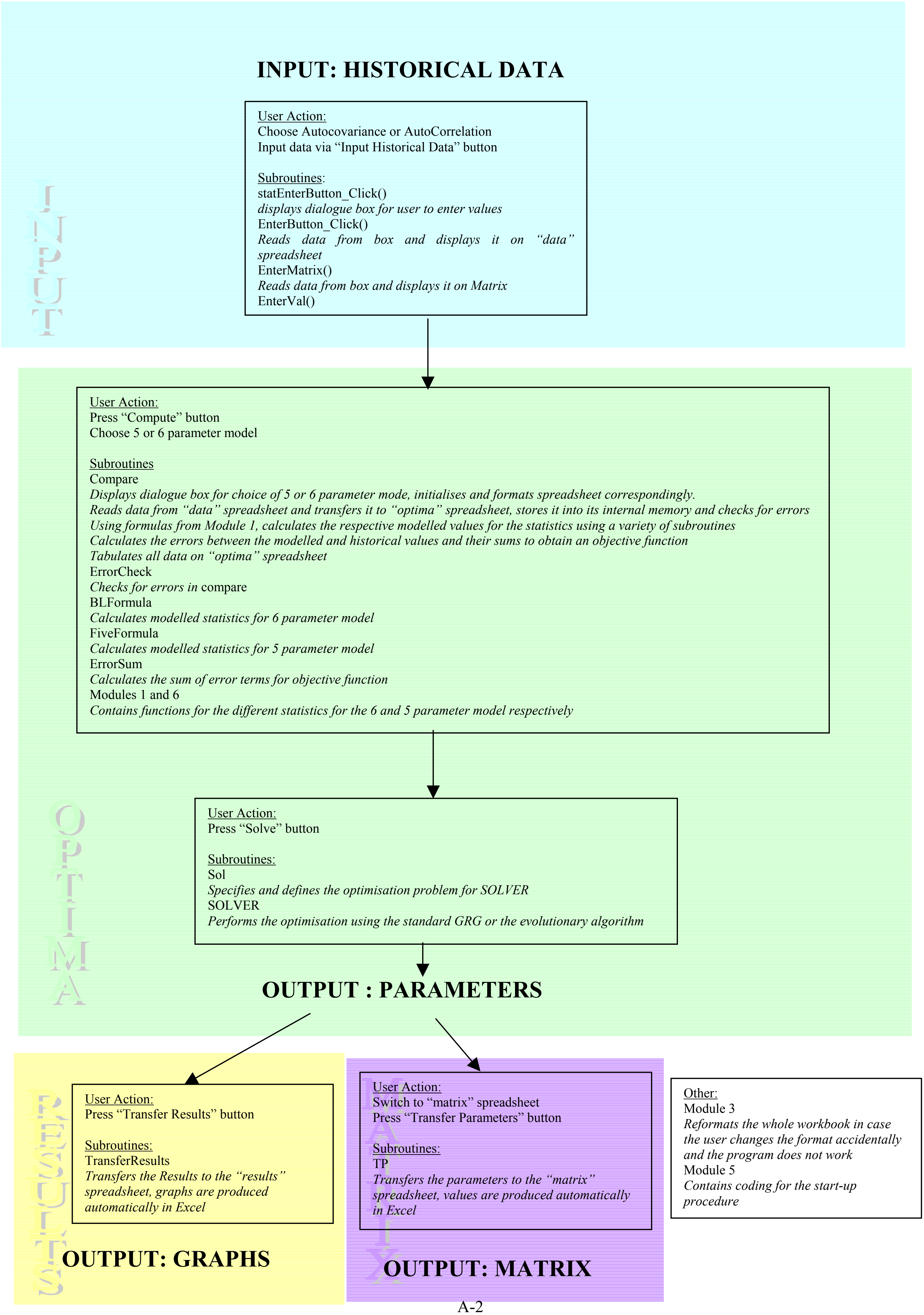
Features

- Written in VBA within a Windows environment
- Incorporated within Excel for easy data manipulations
- Data entry and initial parameters are set via dialogue boxes
- Choice of statistics include mean, variance, autocovariance, autocorrelation, proportion of dry periods and mean duration of dry periods
- Choice of time scales include 1hr, 6hr, 12hr and 24hr
- Choice of use of five or six parameter model
- Sister programme norway was developed for time scales of 1, 2 and 4 days
- SOLVER engine efficiently optimises the objective function using the GRG non-linear optimisation code
- Sister programme evo uses an evolutionary algorithm in its SOLVER engine for the optimisation
- Parameter constraints can be changed easily
- Parameters are tabulated and graphed at user's indication
- Matrices comparing historical and modelled statistics can be drawn up

Coding

Input (required) :	Historical data as entered via the dialogue box
Input(optional) :	Initial values for the parameters Constraints on the parameters
Output :	Parameters from the optimisation
Output (optional) :	Graphs and tables of parameters Matrix comparing modelled and historical data

The code is expressed in the flow chart on the following page. The background is coloured according to the different spreadsheets that the programme is working within. Each text box shows the user action from the user point of view and the corresponding codes and subroutines that are activated. Minor codes (such as the cancellation and deletion of cell contents) have been omitted for clarity.



INPUT: HISTORICAL DATA

User Action:
Choose Autocovariance or AutoCorrelation
Input data via "Input Historical Data" button

Subroutines:
statEnterButton_Click()
displays dialogue box for user to enter values
EnterButton_Click()
Reads data from box and displays it on "data" spreadsheet
EnterMatrix()
Reads data from box and displays it on Matrix
EnterVal()

User Action:
Press "Compute" button
Choose 5 or 6 parameter model

Subroutines
Compare
Displays dialogue box for choice of 5 or 6 parameter mode, initialises and formats spreadsheet correspondingly.
Reads data from "data" spreadsheet and transfers it to "optima" spreadsheet, stores it into its internal memory and checks for errors
Using formulas from Module 1, calculates the respective modelled values for the statistics using a variety of subroutines
Calculates the errors between the modelled and historical values and their sums to obtain an objective function
Tabulates all data on "optima" spreadsheet
ErrorCheck
Checks for errors in compare
BLFormula
Calculates modelled statistics for 6 parameter model
FiveFormula
Calculates modelled statistics for 5 parameter model
ErrorSum
Calculates the sum of error terms for objective function
Modules 1 and 6
Contains functions for the different statistics for the 6 and 5 parameter model respectively

User Action:
Press "Solve" button

Subroutines:
Sol
Specifies and defines the optimisation problem for SOLVER
SOLVER
Performs the optimisation using the standard GRG or the evolutionary algorithm

OUTPUT : PARAMETERS

User Action:
Press "Transfer Results" button

Subroutines:
TransferResults
Transfers the Results to the "results" spreadsheet, graphs are produced automatically in Excel

OUTPUT: GRAPHS

User Action:
Switch to "matrix" spreadsheet
Press "Transfer Parameters" button

Subroutines:
TP
Transfers the parameters to the "matrix" spreadsheet, values are produced automatically in Excel

OUTPUT: MATRIX

Other:
Module 3
Reformats the whole workbook in case the user changes the format accidentally and the program does not work
Module 5
Contains coding for the start-up procedure

Hyetos

The details of this programme are quoted from Koutsoyiannis and Onof (2000).

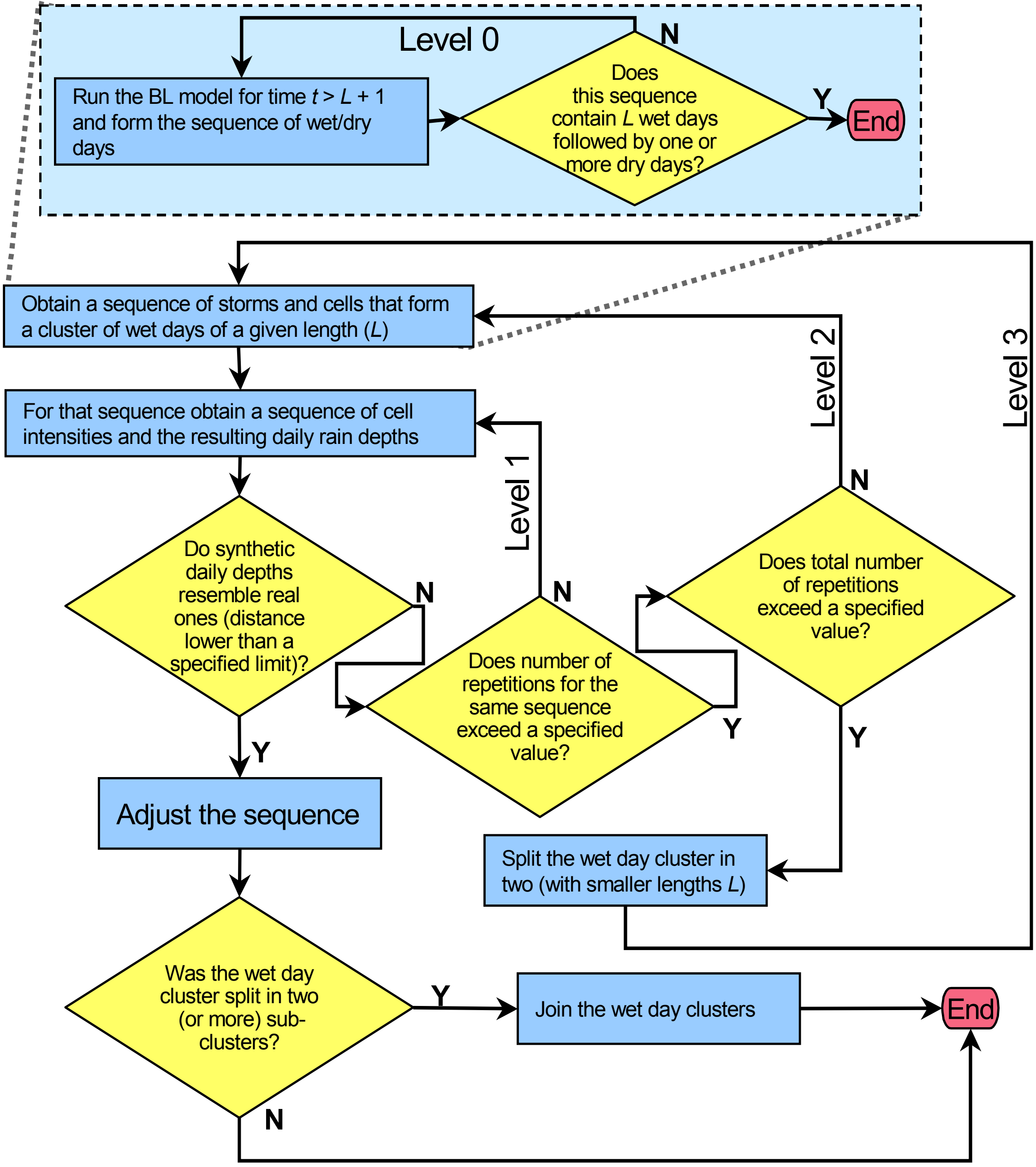
Hyetos uses a windows environment including several forms for input and output of parameters and model options, and graphical forms for plotting and comparing hyetographs and statistics. It includes also help files with instructions and documentation. It can perform in each of the following modes depending on the user selections:

1. Disaggregation test mode (without input). An initial sequence of storms is generated using the Bartlett-Lewis model with the given parameters and then aggregated into hourly and daily scale. The daily sequence serves then as an "original" series, which is disaggregated, thus producing another synthetic hourly series. This mode is appropriate for testing the disaggregation model itself (e.g. by comparing original and disaggregated statistics).
2. Full test mode (with hourly input). In this mode an input file containing hourly historical data must be available. The difference from the Mode 1 is that the original sequence is read from the file rather than generated. This mode is appropriate for testing (e.g. by comparing original and disaggregated statistics) the entire model performance including the appropriateness of the Bartlett-Lewis model and its parameters and the disaggregation model.
3. Operational mode (with daily input). This is similar to Mode 2 the difference being that the input file contains no hourly data but only daily. This is the usual case for the model application. It cannot provide any means for testing.
4. Rainfall model test mode (with hourly input). This is similar to Mode 2 but with synthetic data not disaggregated but generated from the Bartlett-Lewis model with the given parameters. This mode is appropriate for testing whether the Bartlett-Lewis model fits the historical data (in terms of several statistics).
5. Simple rainfall generation mode (without input and without disaggregation). This is similar to Mode 4 but with no input provided (simply the Bartlett-Lewis model parameters are entered). This mode is appropriate for generation of rainfall series using the Bartlett-Lewis model with the given parameters without performing any disaggregation.

Details of the repetition and disaggregation scheme are shown in the figure below, with reference to the disaggregation of daily rainfall depths of a cluster of *L* wet days (preceded and followed by at least one dry day). The scheme was assembled so as to optimise computer time and incorporates four levels of repetition. Initially (Level 0), the Bartlett-Lewis model runs several times until a sequence of exactly *L* wet days is established. Then (Level 1), the intensities of all cells and storms are generated and the resulting daily depths are calculated. These are compared to the original ones by means of the logarithmic distance

$$d = \left[\sum_{i=1}^{L} \ln \left(\frac{Z_i + c}{\tilde{Z}_i + c} \right)^2 \right]^{1/2} \tag{1}$$

where Z_i and $\tilde{Z}_i$ are the original and generated, respectively, daily depths of day *i* of the wet day sequence and *c* a small constant (= 0.1 mm). The logarithmic transformation is done to avoid domination of the very high values and the constant *c* was inserted to avoid domination of the very low values. If the distance *d* is greater than an accepted limit d_a, then we re-generate the intensities of cells (Level 1 repetitions) without modifying the time locations of storms and their cells. If, however, after a long number of Level 1 repetitions, the distance remains higher than the accepted limit, this may mean that the arrangement of storms and cells is not so consistent with the original (and unknown) one. In this case we abandon this arrangement and generate a new one, thus entering Level 2 repetitions. Furthermore, in the case of a very long sequence of wet days it is practically impossible to get a sequence of wet days with departure of daily sum from the given daily rainfall lower than the specified limit. In these cases the sequence is subdivided into sub-sequences (in a random manner), each treated independently from the others (Level 3 repetitions). Eventually, the sequence with distance smaller than the accepted limit is chosen and further processed by determining the lower-level (e.g., hourly) rainfall depths and applying them the proportional adjusting procedure.


Level 0
Run the BL model for time $t > L + 1$ and form the sequence of wet/dry days
Does this sequence contain L wet days followed by one or more dry days?
N
Y
End
Obtain a sequence of storms and cells that form a cluster of wet days of a given length (L)
For that sequence obtain a sequence of cell intensities and the resulting daily rain depths
Do synthetic daily depths resemble real ones (distance lower than a specified limit)?
N
Y
Level 1
Does number of repetitions for the same sequence exceed a specified value?
N
Y
Level 2
Does total number of repetitions exceed a specified value?
N
Y
Level 3
Adjust the sequence
Split the wet day cluster in two (with smaller lengths L)
Was the wet day cluster split in two (or more) sub-clusters?
Y
N
Join the wet day clusters
End

The Evolutionary Algorithm

The field of genetic algorithms is a new and exciting area of research. The idea of using principles from biology in optimisation and search algorithms was initially proposed by John Holland in the 1970's. Holland believed that by applying the same principles in optimisation, he could create an environment where an initially random set of solutions is allowed to evolve into better and better solutions. In this way, problems that were difficult to solve deterministically could be encoded and solved with relative ease. Presently the field of genetic algorithms has itself evolved (!) and now encompasses a wide variety of disciplines, ranging from optimisation to parallel algorithms, and classifier systems to neural networks. It is important to note that genetic algorithms are not the only form of optimisation method based on analogy. Other metaphorical methods, such as neural networks or simulated annealing, are also areas of active research.

Genetic algorithm researchers often have significant differences in their implementation of the algorithm. Therefore, if one were to look within algorithms written by different people, the programme architecture is likely to be significantly different. This is important to note, as there are few strict rules for implementation. The basic genetic algorithm is inherently modular in nature, therefore allowing individuals to add, alter or remove parts without affect the rest of the algorithm. This has led to a wide spectrum of shades of code. This situation is exacerbated by the lack of solid theory (although such research is presently very active) to the metaphorical nature of the algorithm, resulting in all evaluation being based purely on performance. This once again poses problems, due to the random nature of the algorithm. However, the basic principles behind the algorithm do not change, and these are briefly discussed below.

Genetic algorithms can be considered as consisting of two elements. The first is an encoding mechanism which encodes the information from each potential solution set. The very first algorithms used binary forms of code, later methods developed numerical codings. The codes are stored in strings called chromosomes. The second element is a method of evaluating each potential solution and determining its fitness, in other words, determining whether it has reached the objective.

Modern genetic codes usually consist of the following modules: a module to generate a random solution for the population, a module to encode the solutions, a module evaluate the fitness of each solution and finally, a module that applies reproduction and mutation principles to the population. The actual method of implementation is different for each specific algorithm, leading the rich variety of forms the genetic algorithm can take.

The distinguishing feature of a genetic algorithm is the reproduction, or in GA jargon, crossover, processes. These are the procedure by which chromosomes are selected and mixed together to produce offspring, or children. The selection of the parents is usually based on the "Roulette wheel parent selection", where parents are chosen with probabilities according to their fitness. In this way, the most fit solutions will get to reproduce more often, producing offspring of increasingly better fitness. Therefore, the characteristics of dominant solutions, or "super individuals", are quickly spread across the population. The exact process of the crossover is an area of active research. Initial methods used one-point crossover techniques, where the chromosomes were simply cut at one point and the portions switched across. More sophisticated methods are now used.

If crossover were the only procedure to be used, "inbreeding" would quickly arise, such that a single solution would be overly dominant. There would be a lack in the diversity of the population and important solutions may be missed out. Mutation is therefore employed to keep the algorithm in place, continually churning up new solutions, so as to keep the variety within the population at a certain level.

Hybrids of genetic algorithms and normal deterministic methods have been developed. These often prove to be better than the pure GA or optimisation methods when used alone. These hybrid models are of much interest and it is recommended that these models are used if customisation to the BLRPM is to be undertaken.

The robustness of a genetic algorithm has often been a sticking point, since it is often the case that an algorithm may be able to work with one specific problem, but not give as good a performance when transferred to another situation. Genetic algorithm researchers have been searching for an algorithm that is sufficiently robust to work in most situations, but have often failed. Current thinking is that such robustness is not the ideal. Rather, genetic algorithms may be better utilised if customised to the problem, give the solution in an efficient and effective manner.

Other programmes

These computer programmes were minor codes used in specific parts of the project. All the programmes were "bought-in". The programmes are written in FORTRAN, with `gaugestats` and `momentfit` being situated on a UNIX platform. To access these two programs, a telnet connection had to be made to a computer in University College London. For each programme, the input and output are given, with a brief description of its use.

gaugestats

Input: time series data for the historical or a simulated data set
Output: important statistics derived from the time series, such as mean, variance, etc.

`gaugestats` was used to find the major statistics for each month and time scale of the historical data set. The values were then used in the optimisation. It can also be used to find statistics from a simulated time series, although Hyetos immediately does such calculations at the end of the simulation.

momentfit

Input: statistics used to fit parameters to the BLRPM
Output: optimal parameters for the BLRPM

`momentfit` was the initial parameter fitting programme used, before Optima was developed. This programme was used to check the algorithm in Optima.

gev

Input: maximum rainfall depths for time period (years)
Output: modelled Type 1, 2 or 3 distribution for the extreme rainfall depths

`gev` was used to obtain plots for the extreme rainfall depths. Simulated data was analysed and the maximum rainfall depth for each year in the data set was picked out. These values were processed by the programme, modelling a suitable distribution to these values. The Gumbel reduced variate, given in the output, is then plotted with the extreme rainfall depths.

pg/tomo

Input: rainfall time series historical data or simulated time series
Output: power spectrum of the variance with respect to frequency

`pg/tomo` was used to derive the power spectrum for the rainfall time series. The output was a spectrum for each year that was analysed, therefore, the final power spectrum had to be derived by averaging out the values from these years.

Other simple codes were written for the averaging of the power spectrum and the selection of maximum rainfall values from various time scales. These were written in VBA and are very minor codes written to aid calculation and are not explicitly described here.

APPENDIX B

DATA TABLES: CASES 1-4

APPENDIX C

DATA TABLES: Case 4, 8stats and 6stats
for all twelve months

Data tables show for each month the 1-hr modelled statistics for each optimisation scheme over a range of initial values. The corresponding modelled parameters are also included. Some comments are made below each table, drawing attention to anomalies to the general trends.

APPENDIX D

DISAGGREGATION DATA

for various months and optimisation schemes

APPENDIX E

EXTREME VALUE DATA

corresponding to figures 7.1-7.3